MOBILITY 2012

The Second International Conference on Mobile Services, Resources, and Users

ISBN: 978-1-61208-229-5

October 21-26, 2012

Venice, Italy

MOBILITY 2012 Editors

Josef Noll, University of Oslo & Movation, Norway

Alessandro Bazzi, CNR - IEIIT, Italy

MOBILITY 2012

Foreword

The Second International Conference on Mobile Services, Resources, and Users [MOBILITY 2012], held between October 21-26, 2012 in Venice, Italy, continued a series of events dedicated to mobility-at-large, dealing with challenges raised by mobile services and applications considering user, device and service mobility.

Users increasingly rely on devices in different mobile scenarios and situations. "Everything is mobile", and mobility is now ubiquitous. Services are supported in mobile environments, through smart devices and enabling software. While there are well known mobile services, the extension to mobile communities and on-demand mobility requires appropriate mobile radios, middleware and interfacing. Mobility management becomes more complex, but is essential for every business. Mobile wireless communications, including vehicular technologies, bring new requirements for ad hoc networking, topology control and interface standardization.

We take here the opportunity to warmly thank all the members of the MOBILITY 2012 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to MOBILITY 2012. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the MOBILITY 2012 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that MOBILITY 2012 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the areas of mobile services, resources and users.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the charm of Venice, Italy.

MOBILITY Chairs:

Josef Noll, University of Oslo & Movation, Norway
Petre Dini, Concordia University, Canada & IARIA, USA
Pekka Jäppinen. Lappeenranta University of Technology, Finland
Abdulrahman Yarali, Murray State University, USA
Filipe Cabral Pinto, Telecom Inovação S.A., Portugal
Xiang Song, Microsoft, USA
Xun Luo, Qualcomm Inc. - San Diego, USA
Mikko Uitto, VTT Technical Research Centre of Finland, Finland
Sandro Moiron, University of Essex, UK

Mohammad Mushfiqur Chowdhury, University of Oslo, Norway
Masashi Sugano, Osaka Prefecture University, Japan
In-Young Ko, Korea Advanced Institute of Science and Technology (KAIST), Korea
Brent Lagesse, Oak Ridge National Laboratory, USA
Jörn Franke, SAP Research Center - Sophia Antipolis, France
Nils Olav Skeie, University College Telemark, Norway
Gianluca Franchino, CEIICP - Scuola Superiore Sant'Anna - Pisa, Italy
Chunming Rong, University of Stavanger, Norway
Josef Noll, Center for Wireless Innovation, Norway
Aline Carneiro Viana, INRIA Saclay - Ile de France - Orsay, France
Sarfraz Alam, UNIK-University Graduate Center, Norway

MOBILITY 2012

Committee

MOBILITY General Chair

Josef Noll, University of Oslo & Movation, Norway

MOBILITY Advisory Committee

Petre Dini, Concordia University, Canada & IARIA, USA
Pekka Jäppinen, Lappeenranta University of Technology, Finland
Abdulrahman Yarali, Murray State University, USA

MOBILITY Industry Liaison Chairs

Filipe Cabral Pinto, Telecom Inovação S.A., Portugal
Xiang Song, Microsoft, USA
Xun Luo, Qualcomm Inc. - San Diego, USA

MOBILITY Special Area Chairs on Video

Mikko Uitto, VTT Technical Research Centre of Finland, Finland
Sandro Moiron, University of Essex, UK

MOBILITY Special Area Chairs on Mobile Wireless Networks

Mohammad Mushfiqur Chowdhury, University of Oslo, Norway
Masashi Sugano, Osaka Prefecture University, Japan

MOBILITY Special Area Chairs on Mobile Web / Application

In-Young Ko, Korea Advanced Institute of Science and Technology (KAIST), Korea

MOBILITY Special Area Chairs on Context-aware, Media, and Pervasive

Brent Lagesse, Oak Ridge National Laboratory, USA

MOBILITY Special Area Chairs on Mobile Internet of Things and Mobile Collaborations

Jörn Franke, SAP Research Center - Sophia Antipolis, France
Nils Olav Skeie, University College Telemark, Norway

MOBILITY Special Area Chairs on Vehicular Mobility

Gianluca Franchino, CEIICP - Scuola Superiore Sant'Anna - Pisa, Italy

MOBILITY Special Area Chairs on Mobile Cloud Computing

Chunming Rong, University of Stavanger, Norway
Josef Noll, Center for Wireless Innovation, Norway

MOBILITY Publicity Chairs

Aline Carneiro Viana, INRIA Saclay - Ile de France - Orsay, France
Sarfraz Alam, UNIK-University Graduate Center, Norway

MOBILITY 2012 Technical Program Committee

Jemal Abawajy, Deakin University - Geelong, Australia
Ioannis Anagnostopoulos, University of Central Greece, Greece
Payam Barnaghi, University of Surrey, UK
Mostafa Bassiouni, University of Central Florida - Orlando, USA
Paolo Bellavista, University of Bologna, Italy
Rajendra V Boppana, University of Texas - San Antonio, USA
Carlos Carrascosa Casamayor, Universidad Politécnica de Valencia, Spain
Ioannis Christou, Athens Information Technology, Greece
Yan Cimon, FSA/Université Laval - Québec City, Canada
Klaus David, University of Kassel, Germany
Claudia de Andrade Tambascia, CPqD Foudation, Brazil
Amnon Dekel, Hebrew University of Jerusalem. Israel
Raimund Ege, Northern Illinois University, USA
Gianluigi Ferrari, University of Parma, Italy
Randy Fortier, Thompson Rivers University, Canada
Gianluca Franchino, TeCIP - Scuola Superiore Sant'Anna - Pisa, Italy
Xiaoying Gan, Shanghai Jiao Tong University, China
Thierry Gayraud, Université de Toulouse, France
Chris Gniady, University of Arizona, USA
Richard Gunstone, Bournemouth University, UK
Qi Han, Colorado School of Mines, USA
Jiankun Hu, Australian Defence Force Academy - Canberra, Australia
Peizhao Hu, NICTA, Australia
Jin-Hwan Jeong, ETRI (Electronics and Telecommunications Research Institute), Korea
Vana Kalogeraki, Athens University of Economics and Business, Greece
Vasileios Karyotis, National Technical University of Athens (NTUA), Greece
Moritz Kessel, Ludwig-Maximilians-Universität München, Germany
Nikos Komninos, Athens Information Technology - Peania, Greece
Ioannis Krikidis, University of Cyprus, Greece
Abderrahmane Lakas, United Arab Emirates University, United Arab Emirates
Jingli Li, TopWorx, Emerson, USA
Xun Luo, Qualcomm Research Center, USA
Dario Maggiorini, University of Milano, Italy
Barbara M. Masini, CNR - IEIIT, University of Bologna, Italy

Constandinos Mavromoustakis, University of Nicosia, Cyprus
Stefan Michaelis, TU Dortmund University, Germany
Masayuki Murata, Osaka University, Japan
Fatemeh Nikayin, Delft University of Technology, The Netherlands
Shumao Ou, Oxford Brookes University, UK
Knut Øvsthus, Høgskolen i Bergen (HiB), Norway
Evangelos Papapetrou, University of Ioannina, Greece
Marco Picone, University of Parma, Italy
Stefan Poslad, Queen Marry University of London, UK
Daniele Puccinelli, University of Applied Sciences of Southern Switzerland (SUPSI), Switzerland
Daniele Riboni, University of Milano, Italy
Joel Rodriques, University of Beira Interior - Covilhã / Instituto de Telecomunicações, Portugal
Djamel Sadok, Federal University of Pernambuco, Brazil
Ahmed Safwat, Queen's University - Kingston, Canada
Farzad Salim, Queensland University of Technology - Brisbane, Australia
Stefan Schmid, TU-Berlin, Germany
Minho Shin, Myongji Unversity, South Korea
Behrooz Shirazi, Washington State University, USA
Sabrina Sicari, Università degli studi dell'Insubria, Italy
Andrey Somov, CREATE-NET, Italy
Danny Soroker, IBM T.J. Watson Research Center, USA
Tim Strayer, BBN Technologies, USA
Masashi Sugano, Osaka Prefecture University, Japan
Javid Taheri, The University of Sydney, Australia
Miao Wang, Free University Berlin, Germany
Wei Wang, University of Surrey, UK
Rainer Wasinger, The University of Sydney, Australia
Stephen White, University of Huddersfield, UK
M. Howard Williams, Heriot-Watt University, UK
Chansu Yu, Cleveland State University, USA
Ting Zhu, State University of New York, USA

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Mobile Service Business Models for Cities:

A Framework Bridging Public and Business Model Design Parameters

Nils Walravens
IBBT-SMIT
Vrije Universiteit Brussel
Brussels, Belgium
nils.walravens@vub.ac.be

Abstract— **This paper proposes a business model framework that allows the design and analysis of value networks for mobile services in a public context. It starts from a validated business model framework and expands it to include parameters that come into play when a public entity (i.e., a city administration) becomes involved in the value network. In the quickly changing mobile telecommunications industry, this framework offers both an academic and practical tool, enabling the comparison and analysis of complex mobile city service business models that include public actors. After its theoretical development, the framework can be further validated by applying it to specific (inter)national and real-life cases.**

Keywords-Mobile services; business models; public value; public governance

I. INTRODUCTION

The telecommunications industry - and specifically the mobile communications sector - has undergone profound change in recent years, as commercial and public entities aim to find strategic fits while adapting their business models. This also applies to the subsector of mobile service provision (e.g., mobile applications and websites) on a regional, municipal or more local level. New players enter the sector (e.g., Apple and Google), actors shift their business strategies (e.g., Nokia and Microsoft), roles change, different types of platforms emerge and vie for market dominance while technological developments create new threats and opportunities (e.g., NFC and LTE) [10][11][25].

These developments have not been without importance in the context of major metropolitan areas. Both private parties as well as city governments have seen the potential of mobile services, and several, divergent initiatives have been set up and applications or services developed. Mobile services can be particularly attractive in fields such as mobility, cultural activity (discovery), tourism, hotel and catering industry, interactions with government and so on.

However, as these services grow in popularity and importance in the market, questions arise for city governments interested in harnessing the potential of mobile service provision in order to increase the quality of life for citizens in a meaningful way [24]. These questions relate to which roles cities can take up in the value network, how they should interact with emergent players, which data they may leverage in providing services, how they may take up platform roles or how they can create additional public value.

This paper will provide an initial step towards answering these questions by building on the business model matrix, developed and validated in [2]. We will expand it to include business model design parameters that become relevant as soon as a public entity or government actor become involved in the value network. Section II offers a quick reminder of the parameters in the original framework –as they remain important in the newly developed matrix– followed by the development of the additional parameters in Sections III, IV and V. Finally, we propose an expanded framework that can be used both as a design and validation tool in discussing business models for mobile applications, which include public actors. This paper decidedly starts from the perspective of the city and takes mobile services as a case to explore new ways of thinking about business models in a public context and proposes a new theoretical framework to tackle pressing questions in this sector.

II. BUSINESS MODEL MATRIX

In this section, we briefly reiterate the basic concept of the business model framework we will be building on. Ballon [1][2] proposes a matrix that is centered around two types of parameters: control parameters on the one hand and value parameters on the other. It examines four different aspects of business models: (1) the way in which the value network is constructed or how roles and actors are distributed in the value network, (2) the technical architecture, or how technical elements play a role in the value creation process, (3) the financial architecture, or how revenue streams run between actors and the existence of revenue sharing deals, and (4) the value proposition parameters that describe the product or service that is being offered to end users.

For each of these four business model design parameters, three underlying factors are at play, which can be summarized in a dichotomous way, but in reality operate on a scale between the proposed extremes. The use of the matrix as a tool for qualitative analysis has been validated through case studies in several sectors and extensively in relation to mobile services (e.g., in [3]). However, the

 ISBN: 978-1-61208-229-5

specific nature of mobile city services, and more particularly the addition of a public component into the value network, adds increased complexity to the business model. In order to capture the intricacies of combining commercial and public control and value creation, we propose a reorientation and expansion of the business model matrix. This expanded matrix is represented in the figure below and the added parameters will be explained in the following sections.

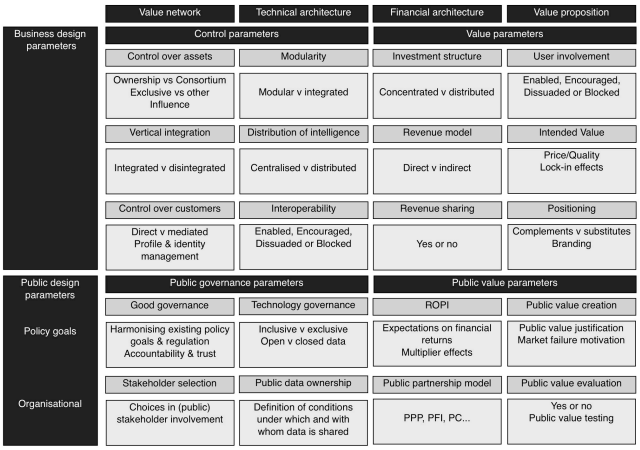


Figure 1. Expanded Business Model Framework.

We note here that all the design parameters important for the business model certainly remain so when a public entity is involved or when certain policy goals are to be achieved. These criteria stay applicable and are not in need of retooling since they were designed with mobile service provision in mind. However, when we take the perspective of a city government or various public bodies, additional business model design parameters become important. We simply refer to these extra parameters as *public design parameters*. In the original business model matrix, a distinction is made between parameters related to control on the one hand and value on the other. This is not different for the public design parameters, however in a public setting we refer to these factors as *public governance parameters* on the one hand and *public value parameters* on the other.

III. PUBLIC GOVERNANCE PARAMETERS

The concept of governance is used in a variety of fields and can be defined in divergent ways (e.g., in strategic management literature [17][29]). This view is however less suited for our approach: the business model matrix assumes a complex value network of several companies, rather than focusing on the internal operations of a single firm. For our purposes, we will use the concept of governance starting from the perspective of the institutions organizing it, i.e., local governments. Our approach is thus based in the idea of *public governance* as described in, e.g., [8]. The United Nations define governance as: *"... the process of decision-making and the process by which decisions are implemented (or not implemented)"* [27] and identifies government as a main actor in governance. It also highlights the added complexity to governance in an urban context, given the large number of actors involved [28]. A policy brief by the Institute on Governance focuses more on the public characteristics of the concept and defines it as being: *"... about how governments and social organizations interact, how they relate to citizens, and how decisions are taken in a complex world. Thus governance is a process whereby societies or organizations make their important decisions, determine whom they involve in the process and how they render account."* [15]. The World Bank [30] offers another take on the process and says governance highlights efficient management of government resources and a mutual respect between citizens and the state.

Depending on the viewpoint, the operationalization of governance can thus be quite variable. For the purposes of defining the governance parameters in relation to the business model matrix, we take note of the UN's definition and can already identify two different layers on which governance can operate, namely in reaching certain policy goals (the implementation process) and organizationally (decision-making). This idea will be expanded upon later on. Elements that are related to the relationships between public and private entities, which stakeholders are involved in the decision-making process, how power and competences are distributed in the value network, the impact of different levels of regulation (transnational, international, national, regional, local), how decisions for or against certain technologies can have effects on the value network and value proposition and so on, are important parameters related to governance, which can be added to the business model through the participation of a public actor. The following section will detail the second set of public design parameters as an addition to the value parameters in the original matrix, namely those related to public value.

IV. PUBLIC VALUE PARAMETERS

The extension of value parameters to public value parameters is a logical one as it is clear that the involvement of a public entity in the value creation and value proposition can have consequences in the public sphere. For example, when public funds are used to develop and deploy a certain service, one might expect a government to justify to tax payers why such an investment is important and whether it fulfills a certain public value.

Mark Moore, author of the seminal work *Creating Public Value* [20], together with John Benington, starts by exposing two ways in which public value can be regarded: firstly, "*what the public values* and secondly, *what adds value to the public sphere"*[5]. He argues that the first question 'what the public values' is a more recent one and can serve as a counterbalance to the top-down determination of what public value *should* be. It empowers citizens to become more active participants in government. However, tensions can form between these two, for example when public service is regulatory in nature (e.g., police) and may impose things on an "*unwilling user*" [20]. With relation to the second question of what adds value to the public sector,

 ISBN: 978-1-61208-229-5

Benington [5] answers with more questions in trying to define what the public sphere or the public itself is, as well as the interesting point on *"what value constitutes in the public sphere, and who decides?"*, exposing questions on power relations, the process of democratic dialogue and absolute and relative values, which are relevant to our analysis. He goes on to detail potential actors that can create value, situate where and how value is created and how it may be measured, and we will come back to this later.

Talbot takes a related approach and identifies different areas in which public values may conflict and proposes that understanding these competing values better, offers a way for public agencies to deal with them [26]. He selects five dimensions on which a public entity should satisfy the public: trust and legitimacy, collectivity, security, personal utility and autonomy. Already, we begin to see similar concepts emerging to the ones appearing in the section on governance and issues such as transparency, responsibility, participation, trust and accountability will be an important part in the further development of the business model matrix.

We take away that the concept of public value is clearly a multi-layered and complex one. For our purposes, we will need to limit the scope in analyzing public value to a more narrow set of parameters. We will define these new parameters in line with the existing business model framework, i.e., per domain (value network, technical architecture, financial architecture and value proposition), but add criteria to reflect the increased complexity when public actors are introduced to the value network.

V. INTRODUCING PUBLIC DESIGN PARAMETERS

The combination of governance parameters and public value parameters to the control and value parameters of the business model matrix, means expanding the framework downward to include additional parameters. The new parameters related to the public domain are explained below. Each time, the first parameter reflects a policy goal, the second an organizational challenge.

A. Governance Parameters Related to the Value Network

1) Good Governance

Similarly to governance, several definitions of what constitutes good governance can be found. The United Nations Development Program states good governance is "*participatory, transparent and accountable*", as well as "*effective and equitable*" and "*promotes the rule of law*" [27]. Hirst [16] proposes a definition, which focuses on the stabilizing elements good governance should entail, and Munshi [21] emphasizes the importance of participation in governing. Graham et al. [15] list five principles for good governance, based on a similar list of eight characteristics of good governance defined by UNESCAP [28], namely *participation, rule of law, transparency, responsiveness, consensus oriented, equity and inclusiveness, effectiveness and efficiency, and accountability*.

Given the relatively vague nature of these concepts and the difficulties in operationalizing them, we will focus on what binds them together: a striving towards equilibrium in governing. This often means finding a balance amongst existing policy goals on the one hand and between those policy goals and existing regulation on the other. As existing policies and regulations can in many cases be contradictory, a striving towards consensus and harmonization of interests is deemed essential in good governance [16]. Since good governance can hardly be regarded as a confined concept [18] and several sources state it should be seen as a process, we propose selecting the trade-offs between often contradictory, existing policy objectives and regulation as an important parameter. In practice, this parameter is dependent on the context in which a certain initiative is taken, but could for example entail an analysis of the goals a service tries to reach and to what extent it contradicts other policies within a government (or e.g., a political coalition) or existing regulation. For example, as more ICT-related regulation comes into play on different decision-making levels (e.g., the Digital Agenda framework laid out by the European Commission [10]), local authorities need to take their compliance with this regulation into account when developing an initiative.

Additionally, we emphasize the concepts of accountability and trust, as it is important to consider which public entity can be held accountable if something should go wrong and how the citizen's rights are protected or can be enforced (see for example [9]).

2) Stakeholder Selection and Management

This organizational parameter refers to the choices that are made related to which stakeholders (be they public, semi-public, non-governmental, private or so on) are involved or invited to participate in the process of bringing a service to end-users (see also the section on governance). In light of the good governance parameter and the striving for balance and consensus described above, including or excluding a particular stakeholder can have consequences for the viability of the final value network and is related to achieving a strategic fit [2] within the business model (cf. supra). Several (sometimes even pragmatic) elements can be important to take into account when deciding on which stakeholders to involve. For example, one aspect could be how competences are distributed among the government actor(s) involved in the value network. When discussing the city, it quickly becomes clear many different levels of government could come into play when offering a certain service, e.g., international, transnational, national, regional, provincial and local. Particularly in the case of large cities or municipalities with large or complex structures, it will be necessary to consider which public organization is responsible for a certain competence or application domain when developing a service, and how these different levels are organized and interact with each other. With the goal of achieving a strategic fit among the actors involved, the

 ISBN: 978-1-61208-229-5

selection process of which stakeholders to involve or not, and how this is decided, is thus important to consider in the analysis.

B. Governance Parameters Related to the Technical Architecture

1) Technology Governance

We borrow this term (more precisely *technological* governance) from [32] who builds upon the concept of *technological citizenship* and links it to how technology is shaped by powerful actors within society. He makes an argument for a more participatory process in which the citizen is the deciding entity in technological choices, which should lead to those technologies "*being more compatible with democratic principles*" than some current "*authoritarian technologies"*. We are not inclined to go as far in this argument, but do recognize the importance of transparency, participation and emancipation in making technological choices, especially by public entities. Choices for a particular technology or platform (e.g., by only offering an iPhone application) may exclude certain parts of the population, something a government should be wary of. This is captured in this parameter through the area of tension "inclusive versus exclusive".

A second element we link to technology governance is the use of open data and whether government information is made available to citizens through the use of ICTs. Many cities and governments are sitting on a wealth of information, which does not find its way to citizens. Okot-Uma [23] lists five important principles related to open government and citizen access to information through digital technologies and ICTs, namely *access, process, awareness, communication and involvement*. Opening up certain data sets and letting developers and the public experiment with them can be an important addition made by a public entity in the mobile services value chain. The choice of a public entity whether or not to open up its data is captured by this parameter.

2) Public Data Ownership

If the decision to open government data to the public is made, the responsible government body should carefully consider the terms under which this data is opened up and to which actors. This is a technological decision in the sense that selecting or limiting the type and amount of formats the data is available in, has consequences to which parties can start working with it (e.g., if the data is machine-readable or not, presented in natural language as well, only available in proprietary formats and so on). Related to this we also consider whether the data is made available to exclusive partners or not and what type of licensing schemes might be in place, as well as their terms. This could be the case when for example a public transportation company decides to provide its real-time travel information to Google, but blocks small developers from accessing the data. These are technical and organizational decisions that can have an important impact on the way the business model is constructed and the final value proposition to the end user.

C. Public Value Parameters Related to the Financial Architecture

1) Return on Public Investment

The phrasing of this parameter is far from new; the notion of expecting a return on public investment in the economic sense is for example mentioned by Margolis [19]. In the context of the business model matrix, we mainly refer to the question whether the expected value generated by a public investment is purely financial, public, direct, indirect or combinations of these, and - with relation to the earlier governance parameters – how a choice is justified. A method, which is often used in this respect, is the calculation of so-called *multiplier effects*, i.e., the secondary effects a government investment or certain policy might have, which are not directly related to the original policy goal. In practice, these effects could be measured by looking at increases in GDP, economic activity, job creation and so on. Calculation of these factors would lead us too far, but we will consider if such indirect return effects are expected or formulated by governments investing in a particular initiative. Also important to consider here is whether these reflections are made *ex-ante* or *ex-post*, i.e., before or after a value proposition is offered to end-users.

2) Public Partnership Model

The organizational parameter to consider in this case is how the financial relationships between the private and public participants in the value network are constructed and under which legal entities they set up cooperation. One example of such a model is the public-private partnership (PPP). Flinders [13] highlights the importance of politics and political tensions behind PPP-constructions as an addition to the traditional analyses from an administrative, managerial, financial or technical viewpoint. While we acknowledge the importance of the political aspect behind PPPs and take into account that political issues may delay or advance particular initiatives, a complete analysis of political tensions underlying certain PPPs is out of scope here. PPPs can also operate in very different areas such as public transport, public utilities, infrastructure and so on. Zhang [31] lists critical success factors for PPPs in infrastructure development such as a favorable investment environment, economic viability, a strong technical consortium, a sound financial package and an appropriate allocation of risk via contractual agreements. Bovaird [6][7] provides an overview of PPP development in the UK and details several potential purposes for and types of PPPs. We also take note of his remarks related to responsibility and risk distribution in this context, namely that the focus should be on the success of the partnership, rather than on that of individual agencies [6].

In the context of the business model matrix, and given the location of the parameter in the financial architecture column, it is clear we choose to emphasize the financial

 ISBN: 978-1-61208-229-5

implications and risk distribution effects of a PPP-model. While other considerations related to the structuring of a PPP are clearly important to the business model design, these are already captured in other public design parameters, e.g., those on *good governance* or *technology governance*. In this perspective, it is also interesting to consider other models, such as a Private Finance Initiative (PFI), a "*more financially-driven PPP, in which the motive for the partnership is fundamentally the readier access to capital finance enjoyed by private sector partners*" [7] or forms of purchasing consortia (PC) which are aimed at seeking economies of scale and bulk purchasing. These and other financial constructions between public and private entities are the subject of this parameter.

D. Public Parameters Related to the Value Proposition

1) Public Value Creation

This parameter examines public value from the perspective of the end user and refers to the justification a government provides in taking the initiative to deliver a specific service, rather than leaving its deployment to the market. A first element that can be of interest is – again borrowed from the broadcasting sector – whether a form of *market failure* is present in a certain domain, i.e., when there is a lacuna in service provision that cannot be met by commercial entities. Of course, depending on the domain, this can be a sensitive discussion (as it is in broadcasting), so together with establishing whether market failure can be identified, we should consider if the fact that there is a specific need in society that is not being met (so that government needs to intervene) is contested by other actors in the value network, or not. And, in the spirit of transparency and good governance, such a justification should also be provided to the public.

We also refer to Moore again here, who, in his Public Value Framework for public organizations [20], proposes some attention points in creating public value (see also [24]): organizational vision (captured in the next parameter by us); strategic goals; links among goals, activities, outputs and outcomes; the range of outcomes; and activities and outputs that create outcomes. We take away here that the goals, outputs and outcomes that public entities wish to achieve need to be clearly outlined and detailed ex ante, so that they can be verified once a service is launched (see the next parameter) and be held accountable (under the good governance principles) should questions on improper behavior arise. The definition of these goals and the promise of their evaluation may also alleviate concerns that can be present with the public.

2) Public Value Evaluation

The organizational parameter we identify as important with regards to how the value proposition is constructed, is whether and how the public value that is (supposedly) created by a public service is evaluated. One way of evaluating the potential success and impact of a public service, can be found in public service broadcasting, with the public value test (PVT) organized by the BBC Trust (the body governing the BBC) and Ofcom (the UK media regulator) as probably one of the most famous examples of such a test. The PVT consists of two parts: the Public Value Assessment (PVA), which is performed by the BBC Trust, and a Market Impact Assessment (MIA), performed by Ofcom. The Trust has a general framework it applies to identify the public value of a service (in some cases ex ante, in others ex post as the debate on which is favorable in which case has not been settled), which is an extension of the public purposes the BBC should fulfill in its role as broadcaster. The parameters of this framework are: reach, quality, impact and cost (and value for money) [4]. These parameters are quite broad and will receive a particular interpretation depending on the service under investigation according to the Trust. The MIA looks at the potential direct and indirect impacts a proposed service may have on consumers and producers of other services in the market [22].

Given the specific nature of broadcasting and the still broad terms describing the PVT, the main take-away towards the business model matrix is whether or not an evaluation is performed in the first place, as well as a description of the form of that evaluation (e.g., a PVT). Clearly, such a test requires clear policy goals that have to be laid out by policy makers and a set of predefined targets such an evaluation should verify.

VI. Business Model Matrix Incorporating Public Design Parameters

These new parameters are important in a context where a public entity becomes part of the value network and have been added to the business model matrix. While this matrix may also be a useful tool in other sectors, we started from its origins in mobile services.

This updated business model matrix, incorporating public parameters, can be used in two ways. In the first place it can guide a qualitative analysis, facilitating the detailed description and comparison of business models in the mobile (public) services industry. By using the parameters to describe different aspects of the business model, a structural comparison between different models becomes possible. Secondly, the matrix can be a useful guide when designing potential business models during the conceptual phase of a service.

VII. Conclusion

This paper set out to build a framework that could facilitate a better insight into mobile service business models when public entities play a role in the value network. We started from the business model matrix, proposed by [1], and expanded it to include public design parameters. Similarly to the distinction [1] makes between control and value parameters, we propose a division between parameters related to governance on the one hand

 ISBN: 978-1-61208-229-5

and public value on the other. Within this division, we delineated eight new parameters to take into account. These operate on two levels: an organizational one, which focuses on how the government organizes itself in realizing the first level, namely the policy goals it sets out to reach. These two levels of analysis are included in the updated matrix.

After making this distinction, we detailed the new parameters and explained their origins. Each of them can be linked up to the original business model matrix, of which the parameters remain applicable. The newly defined governance parameters are good governance, stakeholder management, technology governance and public data ownership. The parameters related to public value are return on public investment, public partnership model, public value creation and public value evaluation. We consider these parameters to be of importance when analyzing a business model in which a public entity (i.e., a city government) is part of the value network.

This expanded framework can be both used as a tool for qualitative analysis (a posteriori) and to design (a priori) the business model of new service initiatives. The parameters allow us to perform a structural analysis of the complex value network of public services and help to identify important aspects that would have been less likely to come to light when only using the business parameters. The addition of the public parameters to the business model matrix adds and interesting and useful layer that allows a more detailed analysis of complex mobile service business models that include public actors.

ACKNOWLEDGMENT

This work was performed in the framework of a Prospective Research for Brussels grant, funded by Innoviris and the Brussels Capital Region.

REFERENCES

[1] P. Ballon, Business Modelling Revisited: The Configuration of Control and Value, Info - The Journal of Policy, Regulation and Strategy for Telecommunications, Information and Media, vol. 9, no. 5, 2007, pp. 6-19.

[2] P. Ballon, Control and Value in Mobile Communications: A political economy of the reconfiguration of business models in the European mobile industry, PhD thesis, Department of Communications, Vrije Universiteit Brussel, 2009, [retrieved: August, 2012] Available online at http://papers.ssrn.com/paper=1331439.

[3] P. Ballon and N. Walravens, Towards a New Typology for Mobile Platforms: Validation Through Case Study Analysis, presented at the 1st Europe, Middle East, North Africa Regional ITS conference (20th European Regional ITS Conference), Manama, Kingdom of Bahrain, 26-28 October 2009.

[4] BBC Trust, Public Value Test: Guidance on the Conduct of the PVT, BBC Trust Regulatory Framework, 2007, [retrieved: August, 2012] Available online: http://www.bbc.co.uk/bbctrust/assets/files/pdf/regulatory_framework/pvt/pvt_guidance.pdf

[5] J. Benington and M. Moore, Public Value: Theory and Practice. London: Palgrave MacMillan, 2011, 314p.

[6] T. Bovaird, Public-private Partnerships: From Contested Concepts to Prevalent Practice, International Review of Administrative Sciences, vol. 70, no. 2, 2004, pp. 199-215.

[7] T. Bovaird, Developing New Forms of Partnership With the 'Market' in the Procurement of Public Services, Public Administration, vol. 84, no. 1, 2006, pp. 81–102.

[8] T. Bovaird and E. Löffler, Public Management and Governance. New York: Routledge, 2009.

[9] H. Davis, Ethics and Standards of Conduct, in Public Management and Governance (T. Bovaird and E. Löffler, Eds.). New York: Routledge, 2009, pp. 311-326.

[10] H. Dediu, The Lives and Deaths of Mobile Platforms, Asymco, 2011, [retrieved: August, 2012] Available online: http://www.asymco.com/2011/02/19/the-lives-and-deaths-of-mobile-platforms/

[11] H. Dediu, The Proliferation of Mobile Platforms, Asymco, 2011, [retrieved: August, 2012] Available online: http://www.asymco.com/2011/09/04/the-proliferation-of-mobile-platforms-continues/

[12] European Commission, Digital Agenda for Europe, Information Society and Digital Agenda Website, 2012, [retrieved: August, 2012] Available online: http://ec.europa.eu/information_society/digital-agenda/index_en.htm

[13] M. Flinders, The Politics of Public-Private Partnerships, The British Journal of Politics and International Relations, vol. 7, no. 2, 2005, pp. 215-239.

[14] R. Freeman and J. McVea, A Stakeholder Approach to Strategic Management, Darden Business School Working Paper N° 01-02, 2001, [retrieved: August, 2012] Available online: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=263511

[15] J. Graham, B. Amos, and T. Plumptre, Principles for Good Governance in the 21st Century, Policy Brief of the Institute on Governance, vol. 15 (August), 2003, [retrieved: August, 2012] Available online: www.iog.ca/publications/policy_briefs

[16] P. Hirst, Democracy and Governance, in Debating Governance: Authority, Steering and Democracy (J. Pierre, ed.). Oxford University Press, 2000.

[17] J. Hunger and T. Wheelen, The Essentials of Strategic Management, Prentice Hall, 2003.

[18] G. Hyden and O. Dele, African Perspective on Governance, Africa World Press, 2000.

[19] J. Margolis, Benefits, External Economies, and the Justification of Public Investment, The Review of Economics and Statistics, vol. 39, no. 3 (August), 1957, pp. 284-291, [retrieved: August, 2012] Available online: http://www.jstor.org/stable/10.2307/1926044

[20] M. Moore, Creating Public Value: Strategic Management in Government, Harvard University Press, 1995.

[21] S. Munshi, Concern for Good Governance in Comparative Perspective, in Good Governance, Democratic Societies and Globalization (S. Munshi and P. Biju, eds.). New Dehli: Sage Publications, 2004.

[22] Ofcom, Methodology for Market Impact Assessments of BBC services, Ofcom, 2007, [retrieved: August, 2012] Available online: http://stakeholders.ofcom.org.uk/market-data-research/other/tv-research/bbc-mias/statement/

[23] R.W. Okot-Uma, Electronic Governance: Re-inventing Good Governance, Commonwealth Secretariat, London, UK, 2000, [retrieved: August, 2012] Available online: http://www.zapataver.gob.mx/work/sites/ELOCAL/resources/LocalContent/1192/9/Okot-Uma.pdf

[24] T. O'Reilly, Government as a Platform, in: Lathrop, D. & L. Ruma (eds) Open Government, O'Reilly Media, 2010, p. 11-30

[25] E. Slivka, Apple and Samsung Claim 99% of Profits Among Top Mobile Phone Vendors, Macrumors, 2012, [retrieved: August, 2012] Available online: http://www.macrumors.com/2012/05/03/apple-and-samsung-claim-99-of-profits-among-top-mobile-phone-vendors/

[26] C. Talbot, Measuring Public Value: A Competing Values Approach, The Work Foundation Research Report, 2008, [retrieved: August, 2012] Available online:

 ISBN: 978-1-61208-229-5

http://www.theworkfoundation.com/Assets/Docs/measuring_PV_final2.pdf

[27] UNDP, Governance for Sustainable Human Development, UNDP Policy Document, 1997, [retrieved: August, 2012] Available online: http://mirror.undp.org/magnet/policy/

[28] UNESCAP, Good Governance, UNESCAP, 2011, [retrieved: August, 2012] Available online: http://www.unescap.org/pdd/prs/ProjectActivities/Ongoing/gg/governance.asp

[29] O. Williamson, Strategy Research: Governance and Competence Perspectives, Strategic Management Journal, vol. 20, 1999, pp. 1087-1108.

[30] World Bank, The State in a Changing World", World Development Report, 1997, Oxford: Oxford University Press.

[31] X. Zhang, Critical Success Factors for Public-Private Partnerships in Infrastructure development, Journal of Construction Engineering, vol. 131, 2005, pp. 3-14.

[32] A. Zimmerman, Toward a More Democratic Ethic of Technological Governance, Science, Technology & Human Values, vol. 20, no. 1 (Winter), 1995, pp. 86-107, [retrieved: August, 2012] Available online: http://www.jstor.org/stable/pdfplus/689882.pdf?acceptTC=true

 ISBN: 978-1-61208-229-5

Mobile Value Chain and Services

The Case of Mobile Donations for Charities

Seyed Mohammad Adeli, Silvia Elaluf Calderwood, Claus Oskar Heintzeler, Javier Huerta, Caroline Legler
Department of Management
London School of Economics and Political Science (LSE)
London, United Kingdom
E-mail: heintzeler@locaid.org.uk

***Abstract*—This case study shows how the use of mobile digital services through smartphones can enhance known value chains of services by increasing the lateral margin value. The particular case discussed relates to a mobile application for charity organizations, i.e., non-profit organizations with the intention of providing help and raising money for those in need. The paper is of relevance for researchers and practitioners, as it demonstrates how computer and business science can be linked to analyze human computer interactions, which may help to solve problems in existing business processes through the use of mobile technology. Based on empirical data gathered from research and interviews during the case study, the paper identifies the most prevalent problems of charity organizations, such as lack of awareness and information, trust, transparency and convenience, and demonstrates how mobile technology can support these deficits in business processes and service value chains.**

Keywords-mobile technology; charity; value chain; value added; location services; mobile design

I. INTRODUCTION

Mobile technologies have been widely studied by both academics and specialists in terms of how their use has changed everyday life in today's society, and also the enterprise relationships between companies and their employees [25]. In recent days, the increase in the availability and popularity of smartphones, like the iPhone (2007) and Android devices (2008), has raised the need to direct attention to re-evaluating the role which mobile devices can play in the delivery of digital services. In this context, mobile devices have changed established value chains and are able to co-create or add value to them.

This paper focuses on analyzing the use of mobile technology associated with smartphones and its ability to add value to the known value chain of charity services. The paper will, firstly, review some of the fundamental theoretical and practical aspects of general value chains for organizations, the value chain for service industries and the value-added possibilities of mobile devices. Next, these established concepts will be customized for the charity service industry on the basis of research and an empirical project, and the development of a smartphone application called "LocAid", which exploits the corners of the charity value chain and creates added value to charities' services.

Finally, the paper will analyze such enhancements in terms of design and provide a set of recommendations which can be used for both defining what value is added to a known value chain when releasing smartphone applications and for the design principles which are required for such development. The paper will finish with the limitations and research directions for further investigation.

II. THEORETICAL BACKGROUND: VALUE CHAIN AND MOBILE TECHNOLOGY

The general concept of the value chain serves as the theoretical model of this paper. In the following, the basic concept will be introduced, its adaptions in the service sector described and the effects of mobile technology identified.

A. The Value Chain

The theoretical model of the "value chain" is first mentioned by Michael Porter [20] in the discipline of strategic management, linking innovation to corporate strategy [17]. It describes how internal activities are developed inside a firm through different steps, which form an economic process, from manufacturing and raw materials to the distribution of the built product.

Porter [20] proposes that a firm can create a cost advantage by reducing the cost of individual value chain activities or by reconfiguring the value chain itself. The concept distinguishes between primary activities and support activities. Primary activities refer to the physical creation of the product, through design, construction, sale and post-sales services such as inbound logistics, operations, outbound logistics, marketing and sales and service. The secondary or support activities help to improve the effectiveness of the primary activities and Porter identifies four main types: procurement, technology development, human resource management and infrastructure [2].

Porter's concept of a value chain is used to model the full range of activities which are required to bring a product or service from conception, through the different phases of production, delivery to final consumers and final disposal after use [12]. The importance of the concept derives from

 ISBN: 978-1-61208-229-5

the fact that it draws attention toward activities which "add value" to the final product or service [12]. It is considered relevant for seeking competitive advantage, reducing costs and identifying ways for differentiation.

Some authors, such as Altenburg [1], argue that the strongest advantage of Porter's model is that it takes into account differences across organizations, suits multifaceted, multidivisional firms and provides information on a firm's strengths and weaknesses. On the other hand, its main limitation is that Porter focuses mainly on products, thereby neglecting services, and only takes into account the internal strategic analysis of an organization, not the external one (industry, customers, etc.), leading to an incomplete analysis of competitive advantage.

B. Value Chain for Services

One limitation of Porter's value chain, as mentioned earlier, is that it does not highlight the importance of exploring new dimensions of the concept, focusing on services, in particular digital services, rather than products. A service approach would give an insight into the flows and transformations by which value is added and might be of great relevance when analyzing service organizations.

One of the main characteristics of services is that their production and consumption happen at the same time. Hence, the service production process itself is the product and, due to the contribution of consumer value, it is more or less co-created. A further characteristic of most services is that, unlike products, services are activities, which are abstract rather than physical and, therefore, are often intangible and impossible to stock. In addition, they are perceived subjectively, making them difficult to evaluate, and factors such as experience, trust, feeling and security play an important role [18]. Based on these characteristics, Gabriel [8] proposes a value chain framework customized for services, as illustrated in Fig. 1.

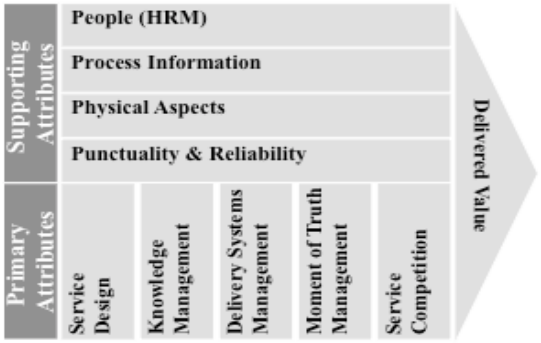


Figure 1. Service value chain, Source: Gabriel [8]

Primary attributes

Service design:

The value of the service needs to be incorporated into the service design. Service designers need to conduct market research and try to be as innovative as possible.

Knowledge management:

Knowledge management refers to the service provider's knowledge about the needs and dynamics of the decision-making process of customers as well as the customers' knowledge about the service.

Delivery systems management:

Services cannot be stored for future use or separated from the provider; this means that services are perishable and inseparable from delivery. Good management of delivery systems increases the convenience for customers and thus improves their perceived value.

Moment of truth management:

The encounter between service provider and customer in the actual moment of delivery of the service can leave a positive or negative impression in the customer's mind. It can build or destroy trust and confidence and can dictate buying decisions for the future.

Service competition management:

Customers have a choice between different competitors. Therefore, providers need to stimulate their clients even after the service delivery. Efficient after-sales management and a high quality of service can increase the perceived value of the service.

Supporting attributes

People:

People are important in the co-creation of value due to the simultaneous use and production of services. Customer expectations need to be matched with offered service to prevent a perception gap.

Process information:

The service provider and their employees need to be aware of their service processes and the generation and delivery of the service value. Transparency and the availability of information, through, e.g., IT, are of great importance in this step.

Physical aspects:

The physical aspects refer mainly to customer services and also include tangible aspects like the office's appearance. Accompanying the customer in his preferred way throughout all primary activities and signaling the

 ISBN: 978-1-61208-229-5

value of the service throughout this process is, therefore, crucial.

Punctuality and reliability:

The time aspect is significant in the service industry and relates directly to the service quality. Reliability implies a level of consistency and assurance for the customer.

In a similar framework, Nooteboom [18] attempts to develop a generalization of Porter's framework, corresponding to different types of service industry. This industry differentiation, based on the central features of the value adding process, is believed to enable easier identification of sources of inefficiency, to detect opportunities for value added and to be crucial for increasing transparency.

The research question this paper seeks to answer is: how can mobile applications, through carefully crafted feature design, enhance different steps within the service value chain? The model described above is used as a theoretical framework for the purpose of this study, since it offers a more viable perspective than Porter's original framework.

C. Value-Added through Mobile Technology

Mobile technologies and, more specifically, mobile applications have unique attributes which can add significant value to a company's service value chain. The literature identifies, in particular, three features as fundamental supporters in today's business:

Connectivity

Connectivity or mobility refers to the interdependence of time and place. A wireless infrastructure offers "anytime, anywhere" communication and information exchange [4]. It is especially valuable for time-critical or spontaneous needs [14] and it is useful to employees and customers alike in that mobile services provide both user groups with easy access to the most up-to-date information [3, 13].

Personalization

Mobile devices are typically assigned to single users, who can then personalize the interface and application settings of the devices [4]. Especially for interactive and dynamic mobile services, personalization or customization is fundamental for supporting user satisfaction and the efficiency of a system, according to Barnes [3] and Coursaris et al. [4]. Moreover, mobile technologies support an easy modification of content, the repetitive and simultaneous consumption of information by different users and fast and cheap reproduction [3].

Localization

The Internet has the ability to localize specific places (e.g., IP address). Mobile technologies can extend this localization feature by also localizing users (e.g., a mobile worker) and items (e.g., tracking a shipment) [4]. This feature is strongly demanded, especially with respect to today's development of mobile applications.

The identified attributes can be very valuable throughout different stages of the service value chain. They can play a significant role in service design, knowledge management and delivery system management. Moreover, mobile technologies are able to assist all supporting attributes (people, process information, physical aspects, punctuality and reliability) of the value chain.

However, while improving the connection between the customer and the company, some problem areas may arise. As Gabriel [8] argues "the more convenient the system, the better the perceived value by customers". This points toward the need to give crucial attention to ease of use and the perceived usefulness of the mobile device in order to ensure that customers do actually use the device, that is that they engage in the "cognitive effort" [4, 13]. Moreover, privacy and safety in information exchange [4] are often perceived as risks in mobile services. In particular, customers can lack trust in monetary transactions in mobile commerce and, therefore, these should be given special focus in the service value chain.

III. MOBILE VALUE SERVICE FOR CHARITIES: "LOCAID"

In order to illustrate and understand how innovative mobile services might add value to established value chains, this section studies the case of the mobile application "LocAid" in the context of the charity industry. It shows how LocAid's specific design features, identified in market research and in interviews with charities and the charitable society, can add value to the value chain of charity services. Firstly, the general process and work of charities will be described and a framework for a charity value chain proposed. Following this, the LocAid project itself will be introduced and its effects on the value chain illustrated.

A. Charities and their Value Chain

A charity organization can be defined as a non-profit organization with the intention of providing help and raising money for those in need. According to the Charities Act of 2006 [26], charitable activities include, among others, support for health care, poverty prevention, community development and environmental issues. These activities range from a local to an international level. To finance their work, charities rely mainly on external funding. Individual donors constitute the main source of income, followed by charitable trust grants, fundraising initiatives, asset investments, trading subsidiaries and charity shops [26].

The general relationship between charities and the beneficent public, and the basic charity operations required to pass public resources to those in need, is illustrated in Fig. 2.

 ISBN: 978-1-61208-229-5

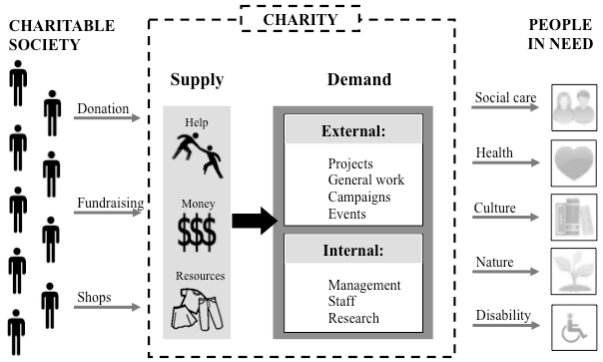


Figure 2. Charity process, Source: own illustration

This shows that charities may receive financial, human or physical resources from charitable citizens through a range of activities. These activities can be broadly categorized into donating, fundraising and giving to/buying from charity shops. The overall supply of resources is then used to support the external and internal demands of charities, such as the funding of specific campaigns and the management of the organization, in order eventually to help people in need.

The activities of donation, fundraising and charity shop use, with which people can engage, are rather diverse. Fig. 3 depicts the different ways of contributing within each category.

The first activity, **donating**, is the process of giving money to a specific need. While charitable people can donate on a one-time or regular basis (single/regular donation), they can also make donations on behalf of somebody else (gift donation) or based on their own will (legacy).

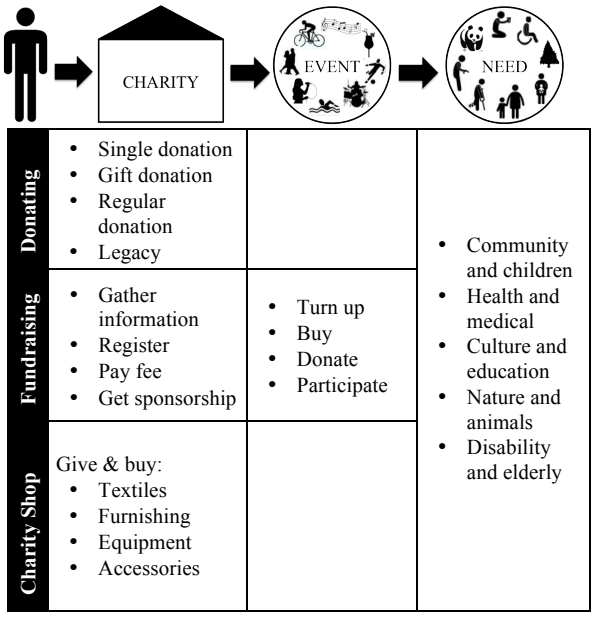


Figure 3. Activities with which charitable people can engage, Source: own illustration

The research for the LocAid project revealed that most charities do not specifically state to whom/which organization their donations will be given. Many charity websites did not show which particular projects they operated and remained relatively inexplicit about the general work they do. This lack in transparency might discourage donors from becoming involved in charitable giving and reduce the trust in charity organizations.

The second activity, **fundraising**, can be defined as the process of giving time and effort to a specific need. Potential participants may need to gather information on event details, register for an event, pay a participation fee and collect sponsorships from other people. Moreover, the nature of the event and the degree of involvement influence the fundraiser's activity. Fundraisers may just turn up to support others mentally (e.g., cheering at a marathon), buy an event ticket or items at the event (e.g., registering for a party or buying a cake), donate at an event (e.g., donating at a gala dinner) or actively take part in the event (e.g., running a marathon).

Internet research for fundraising events showed that charities do list their own events on their websites. However, there are only a few websites which list collectively the events of various charities in a specific area. Hence, a higher level of participation might be achieved with a clear overview of such events.

Finally, people can donate their resources or money to **charity shops** by bringing their own goods or purchasing second-hand items. The traded items can be textiles (e.g., clothes, shoes), furniture (e.g., mirrors, photo frames), equipment (e.g., sport equipment, books, CDs) or accessories (e.g., bags, jewelry).

The project's research revealed that the number of charity shops is growing, especially in Western countries. This might be due to the throwaway culture which has emerged over the last few decades, but also to the current recession, which makes people feel less able to give money yet perhaps still able to donate unused items. Moreover, a social trend was observed, whereby people would like to do something "good" while buying something. Campaigns such as Fair Trade might confirm this trend.

Charity Value Chain

In contrast to services in commercial sectors, the charity sector is strongly driven by the beliefs of people who want to support a specific cause [15, 23]. It is crucial that charities understand why people give to causes and communicate their services accordingly to achieve a long-term commitment [9]. Based on LocAid's market research and interviews, the most prevalent challenges for charities are the lack of information made available, and their trust and transparency, both of which are instrumental in deterring people from giving more to charities. In addition, some charity services, especially through the Internet, are inconvenient to use. Hence, this paper will draw attention to the information, linked to general charity awareness, trust,

 ISBN: 978-1-61208-229-5

transparency and convenience, and how these issues could be mitigated through any kind of value added within the value chain.

According to Saxton [23], people can be motivated to give charitably on different levels, from a shared identity ("I share their vision") to the effects on their local environment ("It makes a difference to me"). Other scholars identify, therefore, the distinct importance of brand management for charities in order to communicate and symbolize the specific beliefs of charitable people, motivate them and facilitate the process [10, 11]. Hakinson [10] divides a brand into functional attributes (the cause) and symbolic values (brand values) like humanity, impartiality, neutrality or independence. Her research shows that charity managers use brands to fulfill a range of organizational objectives such as raising awareness, building trust, fundraising, educating or lobbying. The small amount of existing research on charity organizations and their processes shows that, in order to create value, a distinct focus on the cause and its symbolic values is required.

The framework in Fig. 4 is an attempt to identify potential aspects of a charity value chain as a service. It is based upon the service framework of Gabriel [8] with some adjustments taken from market research and interviews with respect to the LocAid project.

Primary attributes

Service design:

The design of a charity service will be oriented towards beneficent people and their specific motivations for a cause. Marketing might play an important role in incorporating the cause, the charity value and the resource provision into the service design or even in building a specific charity brand. Customer segments, and specifically their intrinsic motivations, might be identified through market research to enable an effective service design.

Knowledge management:

The knowledge management phase could be a potential step in enabling effective information provision about donors and their profiles. Customer data would need to be stored intelligently in order to match the specific needs of customers with identified relevant causes and projects. The organization would also need to ensure that beneficent people are aware and sufficiently informed about the charity and its service value.

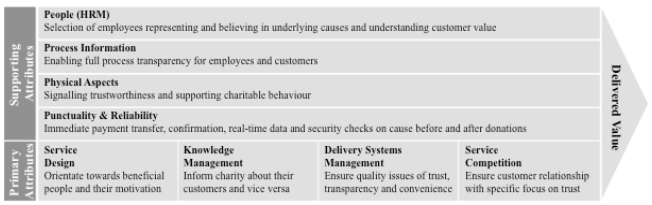


Figure 4. The charity value chain, Source: own illustration

Communication and feedback processes through customer service might strongly support the effectiveness of the knowledge management phase.

Delivery systems management:

The delivery phase of a charity service would aim to ensure that the most prevalent challenges of trust, transparency and convenience are addressed. Specific focus might be given to the convenience of the search, selection, payment and registration processes for a cause. This might be equally important for all service channels, whether on the web, via a mobile, via a call-center or by personal interaction. Trust and transparency might be enhanced within this step through, e.g., successful fundraising events, a strong focus on payment security, trust seals or an immediate donation confirmation.

Service competition management:

Strong competition for donations has been observed during LocAid's market research. The service competition management phase would be a potential value chain step to signal the positive difference of the charity in comparison to its rivals in the market. As charities try to incentivize customers to donate on a regular basis, long-term satisfaction will be crucial for charities. A focus on trust, individual needs and the visible effects of donations might be supportive in building strong customer relationships. Communication after donations might be targeted, e.g., through regular updates on a cause.

Supporting attributes

People:

The project's research also revealed that the value of charity services was extremely dependent on value co-creation with customers since, without the support of charitable people through help, money or resources, a charity itself would be meaningless. Hence, donors should feel their own importance throughout the whole value chain. People might also refer to the employees of the charity, who should represent and believe in the underlying causes. They should signal seriousness, generosity, sensitivity and customer-friendliness and try to build trust in order to match the offered service with the donor's expectations.

Process information:

Throughout the value chain, the charity process would need to be as transparent as possible. Charitable people should be able to know the destination of their contribution and its effects on a specific cause. It is proposed that employees should be able to access this information and provide it, if appropriate, to customers. Optimized information technology might help in this step to ensure data quality, tracking and provision.

 ISBN: 978-1-61208-229-5

Physical aspects:

The physical aspects of charities would signal the service value with an emphasis on trustworthiness and the charitable behavior of giving people. These aspects might be divided into "Marketing" - online (e.g., website, application) or offline (e.g., catalogue, flyer) - and "Facilities" (e.g., office, furnishings, charity shop).

Punctuality and reliability:

Reliability, and punctuality which is closely related, would be crucial for charities to build trust. Customers would need to be assured that the charity service is serious and reliable, for instance in providing the donated money to the corresponding cause efficiently. Immediate payment transfer and confirmations, real-time data, regular data check-ups and security check-ups on causes before and after donations might be able to support this value chain step.

B. The Local Aid Project

The mobile application LocAid (i.e., "Local Aid") was developed in the context of a university project at the London School of Economics and Political Science. It was designed according to the findings of market research, interviews with charities and charitable people and the value chain identified above. In the following, the application itself and a brief overview of the project are presented.

The application LocAid is a mobile application that allows users to find, support and connect to local charity organizations. The application offers the three main functionalities of donating to local charity projects, registering for local fundraising events and finding local charity shops. The innovations put forth through LocAid are driven primarily by the three distinct characteristics of high transparency, local applicability and mobility. The idea is based on the concept of offering value added for users and charities through enhanced information provision and local charity awareness to increase local charitable giving.

The project of the application development was organized into two interrelated, parallel work streams, one focusing on the foundation, justification and evaluation, and the other on the development of the application.

In the first stream, a market analysis, surveys and interviews with charitable people and charity organizations were performed to obtain an understanding of the charity market and its processes and needs and to identify the concrete definition of the application and its required functionalities and design. The surveys were also used to justify different aspects of the initial requirements. The participants were selected as a potential user audience, in order to evaluate the importance of different features of the application from their perspective. The survey feedback was found to be relatively positive, with 96 percent of people considering the application to be useful and more than 60 percent strongly agreeing that it would encourage them to engage more in charitable activities. Users also indicated great interest in additional functionalities such as a map, calendar and news features. The most prevalent issue for participants (more than 70 percent) was the security of the payment method, which was viewed as critical in building trust before using the application. Two focus groups were employed to evaluate the innovation, its usefulness, and potential, additional functional requirements. To evaluate LocAid from a business perspective, a business model was created with a specific focus on user development, the revenue model and the cost structure.

The second stream, the development process, consisted of four main steps. It was based on the waterfall development model [22], allowing iterations between all process steps and using unified modeling language (UML) to complement the design process. The implementation was carried out in two steps: firstly, a functioning GUI prototype was created and feedback was received through the focus groups for further improvements; secondly, the development of a rudimentary real prototype was started in XCode. Issues during implementation were related mostly to the availability of data on charity projects. Most of the available data were only for the charity itself as a whole, but not for specific projects, as the application requires. Future steps, therefore, would include considering direct cooperation with a larger charity with strong local involvement and highly accessible project data. This paper focuses mainly on the computer–human interaction, rather than the implementation issues.

C. Value Added Mobile Application

As described in Section III A, the literature has identified brand and belief creation as two ways to support charities and their value creation. Mobile technology might be another recent way to add value to the charity value chain. Through its distinct characteristics, localization, personalization and connectivity, it might be able to tackle the most prevalent problems of charities which deter people from engaging more with them, in particular the lack of awareness and information, trust, transparency and convenience. The mobile application LocAid was designed to address these problems and is an exemple of an innovative mobile technology with great potential to add value to the value chain of charity services.

Awareness and Information:

During the "service design" and "knowledge management" phases, sufficient awareness and information were regarded as being fundamental to market charities' projects, events and general work. Commonly, charities send street volunteers to inform people, start campaigns to increase awareness of specific projects and have a website presence to keep users up-to-date about events and their work. However, these activities reveal difficulties in explicitly targeting the charitable people and in addressing their individual information needs.

Mobile technologies might be able to address these difficulties through their localization, personalization and

 ISBN: 978-1-61208-229-5

connectivity characteristics. In the case of LocAid, functionalities were incorporated to locate charitable individuals and to show specific donation projects, fundraising events and charity shops "around" them, thus customizing the application to individual needs. Moreover, the application was designed in such a way that personal accounts were offered in order to tailor the content to the specific user (e.g., users get an overview of their past donations or events and receive updates on ongoing projects). Finally, LocAid was provided with a feature to give users information when and where needed, making charitable giving a real-time activity and keeping the individual user informed at any time about the current status of their beneficial actions.

Trust:

Trust was identified as a critical concern throughout the whole charity value chain and generally as a complex, prevalent factor in every financial transaction. Charities should build relationships with a charitable society and increase their involvement to gain and sustain the public's trust. Traditional advertising channels, such as physically approaching people, attempt to develop trust and customer relationships through personal contact. However, many people feel pressured and hence refuse to become involved in this direct approach.

Mobile technologies might create a two-way connection between the charity organizations and the users without pressurizing them. Moreover, customer relationships might be built through personalization. LocAid, for example, includes features to display updates of projects to which users have donated and the most popular projects of other users. In addition, the local focus of the application was chosen to address the trust issue, as local charities might often be better known and their projects can be visited in person. Finally, during the development of LocAid, a networking functionality was considered, connecting charitable people through the application, creating a community and thus eventually developing a lock-in effect.

Transparency:

Transparency was found strongly to influence trust and was seen as crucial for supporting activities in the value chain, such as "process information" and "systems delivery management". The surveys and interviews showed that not only the process of money transactions, but also the money's destination and effect should be transparent. Charitable people were critical of the fact that they do not know where their money goes and their perception of charities' transparency was very often low. Almost all reported that their most important concern was to see the actual result of their charitable actions, leaving them with the desired satisfactory feeling of having done something good.

Mobile technologies might increase transparency through local applicability and customized content. LocAid was designed to give attention to local projects and events in order to increase the perceived visibility of donated money and its effect on local causes. Accordingly, beneficent people can help causes where they see the actual results, in contrast to foreign aid support, where users often feel wary about the destination and use of their money. In addition, personalized features, such as receiving feedback and updates on projects to which a user has donated money, were chosen to foster customer relations management and, thereby, to increase transparency.

Convenience:

Convenience is a factor which has received more attention in recent years due to the time constraints in today's society. As market research and surveys show, it has developed into a core focus in the charity value chain process, especially during the "service delivery" phase. Users engage with a service if it is simple, convenient and efficient. Conventional efforts to offer convenience to charitable people, such as actively approaching people on the street rather than asking them to go onto websites and visit charity offices, or sending forms for event registration via email, cannot meet individual needs to engage in charitable activities at the right time and in the right place.

In constrast, mobile technology can offer service "around the clock", giving the advantage of serving customers whenever and wherever it is convenient for them.

Furthermore, as previously indicated, LocAid's functionality design was focused on personalization and localization (e.g., giving reminders of upcoming events for which users have registered or simple directions to charity shops near the user), linking charitable giving with a comfortable service provision. Related to this, simplicity in design was seen as fundamental to providing convenience, leading to strict guidelines during the development (e.g., the steps required to carry out a donation or fundraising registration should not exceed 3-4 clicks).

The above customer-focused discussion demonstrates the potential ability of LocAid to add value to the existing value chain of charities. Aside from acquiring, serving and satisfying charitable people in a more efficient and effective way, the application may also support charities in their information management and operational efficiency. For example, donations and fundraising registrations can be tracked in real time, new information can be communicated instantaneously and marketing can be conducted through a more targeted approach – all offering the potential for competitive advantages in services and processes. Accordingly, LocAid might not only help charities to deliver a better service, but simultaneously offer benefits to charities, the benevolent society and the people in need, hence acting as an intermediary co-creating value between the three interrelated parties.

 ISBN: 978-1-61208-229-5

Overall, the case of the development of LocAid demonstrates how the effective design of mobile technologies might be able to address prevalent problems and add value in established value chains. The characteristics of localization, customization and mobility were systematically applied to the design of the application in order to fulfill its value added need.

IV. DISCUSSION

The LocAid case shows how the characteristics of mobile technology can add value within an established industry, more specifically within the charity value chain. However, some limitations of the framework and the mobile technology effects need to be considered. Firstly, the relationship of the charity value framework with the literature will be discussed and, secondly, the issues of the mobile technology effects will be described.

The framework is strongly related to Gabriel´s [8] proposed service chain framework but, unlike in the original model, the activities of "moment of truth" and "delivery system management" are combined. For charities, these two activities cannot be differentiated as the actual "service moment" of a charity often cannot be defined due to its subjective nature. People will define their service moment differently: for some, the payment to a cause will be the main service moment, while for others it will be the actual resource provision or positive effect in the future. The main relation to Porter´s original model is the differentiation between primary and supporting activities and the fundamental idea of how value is created within a "chain". The charity value chain is, in contrast to Porter's original model, a service value chain, which emphasizes not the creation of a product but the co-creation of value with its customer.

The charity value described in this paper is formed mainly from market research work (interviews). The academic literature was not found to be sufficiently detailed and was too generic in some cases to be logically conclusive. This is a shortcoming of this work, as the model might require further testing. In addition, the model framework proposed for the charity value chain is only validated for the case of London, or at most the UK; hence, attempts to extend the results to other contexts would require a reassessment of the assumptions for calibration.

The positive effects of the mobile technology on the charity value chain have certain problematic characteristics, which will be critically discussed for each attribute. An issue for all attributes is that the value added can usually only be leveraged if the charity fulfills certain prerequisites (e.g., transparency can hardly be enhanced if the charity does not provide sufficient data on its processes). This issue is strongly related to the key implementation issue regarding available data.

Considering the **supporting attributes** of the value chain framework, *people* and *physical* attributes are unlikely to be influenced by mobile technology. *Process* transparency and increasing trust through *punctuality and reliability* attributes can be improved, but only if the mentioned complementary attributes are present (e.g., the charity needs to be reliable before a mobile technology can add value). Regarding **primary attributes,** mobile applications add value to the *service design* and specifically to general charity awareness. Charities need to be aware that, in the case of LocAid, these positive effects can occur similarly for all cooperating partners of the mobile application service provider. Consequently, the service could be used more out of a necessity to compete rather than as an idea to gain any value added. The *knowledge management*, *delivery systems management* and *service competition* can be affected very positively through the distinct characteristics of personalization, localization and connectivity of a mobile application. Similarly to the supporting attributes, significant value added can only be fostered if the required conditions are fulfilled.

Further limitations of value added through mobile technology arise from the effects on trust and transparency, in combination with taking payments through a mobile application. Trust is a complex concept and a prevalent and important factor for every financial transaction. The ability to measure trust is limited by the fact that it is a multidimensional socio-technical factor which may be differently interpreted by every individual and which has received numerous different definitions [5, 7, 19]. Most scholars agree that trust is a belief in "favorable expectations" [5] based on previous interactions. The problem is that a mobile intermediary increases the number of parties which need to be trusted, in this case not only the charity itself but also the mobile service, leading very often to an as yet unaddressed problem of perceived security. Trust and security, if not perceived by a user, have been identified as major inhibiting factors in user acceptance of payments through a mobile application [5, 16]. Security can generally be divided into objective and subjective security**.** Objective security denotes the concrete technical details which are unlikely to be perceived by the consumer. Subjective security is the perception of a user that the mobile payment procedure is secure and can be seen to combat the perceived risk [5, 6, 21, 24]. Consumers often perceive payment solutions as insecure, thus do not trust them and are therefore unwilling to use them.

The positive effects of mobile technology could be further mitigated by personal characteristics of charitable people, such as their age, beliefs or values. Mobile technology and especially payments through mobile applications are used mostly by younger generations. Because the most charitable group of people is aged between 45 and 64 years [26], their adoption, or even knowledge, of mobile technology can often be limited. Furthermore, the local aspect of LocAid, based mainly upon the localization feature, could go against the beliefs of many charitable people, who generally come from developed countries and often see no reason to donate or support local

 ISBN: 978-1-61208-229-5

charities but want to help foreign poorer developing countries. Ultimately, the idea of an extra service fee due to an additional intermediary could put many people off because, firstly, the donated money could be reduced and, secondly, some believe a charitable intermediary should not aim to gain any benefit at all. This concern should be taken into account in any business-model development for a mobile value service within a charity value chain, for instance by not charging donors at all and charging charities only to the extent that the value added exceeds the additional service charge.

Finally, the effects of the LocAid case need to be critically debated from the perspective of the overall charity industry. Firstly, even though the localization feature can indeed add great value if a charity supports local projects, charities with non-local projects or no possibility to provide local individual information have only limited or no use for the mobile value service. The distinct localization feature, therefore, only applies to charities with local projects. Secondly, the application itself is limited within the charity industry because it does not consider volunteering services, which are of great importance to many charities. The volunteering process often involves a higher level of commitment, and specific skills and training and differentiates itself from donations, fundraising and shop functionalities for any application development.

The LocAid case shows how design specifications can be derived by analyzing the specific value added of the application in relation to the industry into which it is introduced. Alongside innovation, the application was designed in order to signal quality and generosity to overcome trust constraints, but also to incentivize users (e.g., the color green was chosen as the main color due to its signaling of generosity, support and money). Developers and graphic designers should work hand in hand to produce a coherent design which suits the specific requirements of an industry.

The discussion shows that the proposed charity value chain and the effects of the mobile technology and its value service can generally lead to value added, but both the framework and the value added are limited due to the framework's uncertainty, the intermediary character of the mobile technology and the general trust issue within the charity industry.

V. CONCLUSION AND FUTURE WORK

The approach to understanding the role that mobile applications such as LocAid can have in value chains is an area which has not been researched in depth and companies have been slow to understand and plan for future implementations. The waterfall model for the design of mobile applications, when used with sufficient care and vision, is still adequate for providing solutions when required.

The proposed charity value chain framework shows how value is co-created with the customer and which specific attributes can add value to this service. The specific issues of trust, transparency and convenience in the charity sector offer a basis for analyzing the positive effects of mobile technology. The distinct mobile technology features of localization, connectivity and personalization can be related to each value chain attribute and offer strong value added overall.

In terms of design, the carefully crafted attention to detail, in terms of application design, services, trust, etc., allows the provision of an integral solution for the delivery of this type of service, which has been positively embraced by practitioners in interviews with charities in London, and there is interest in releasing the application and its future enhancements in the real life market.

This research paper contributes to the field by presenting a new, business-oriented direction for research in computer science. By focusing on human-computer interactions in relation to specific value chains, it encourages academics and practitioners to work together in order to achieve mutual benefits. In addition, the very practical findings of this paper can help established services to understand the value which new technologies, in particular mobile technology, can give to their businesses and to create an interest in innovations and new developments in the future.

Future research should further assess the proposed value chain framework but also try to identify more specific features of mobile technology which can create value added, and show how practitioners in related industries and developers can use these opportunities to devise practical guidelines such as design specifications.

REFERENCES

[1] T. Altenburg, "Donor approaches to supporting pro-poor value chains," Report prepared for the Donor Committee for Enterprise Development Working Group on Linkages and Valu Chains, Bonn, Germany, July 2006.

[2] P. Ballon, "Control and value in mobile communications: a political economy of the reconfiguration of business models in the European mobile industry," unpublished.

[3] S. Barnes, "The Mobile Commerce Value Chain: Analysis and Future Developments," International Journal of Information Management, vol. 22(2), 2002, pp. 91–108, doi:10.1016/S0268-4012(01)00047-0.

[4] C. Coursaris, K. Hassanein and M. Head, "Mobile Technology and the Value Chain: Participants, Activities and Value Creation," International Journal of Business Science and Applied Management, vol. 3(3), 2008, pp. 14–30.

[5] U. Cyril, G. Gan, J. Ademu and S. Tealla, "Modelling User Trust and Mobile Payment Adoption: A Conceptual Framework," Communications of the IBIMA, vol. 3(29), 2008, pp. 224–231.

[6] T. Dahlberg, N. Mallat, J. Ondrus and A. Zmijewska, "Past, Present and Future of Mobile Payments Research: A Literature Review," Electronic Commerce Research and Applications, vol. 7(2), 2007, pp. 65–181, doi:10.1016/j.elerap.2007.02.001.

[7] T. Dahlberg, N. Mallat and A. Öörni, "Trust Enhanced Technology Acceptance Model – Consumer Acceptance of Mobile Payment Solutions," Proc. 2nd Mobility Roundtable, Stockholm, Sweden, May 2003.

 ISBN: 978-1-61208-229-5

[8] E. Gabriel, “Value Chain for Services: A New Dimension of ‘Porter’s Value Chain’,” IMS International Journal, vol. 34, 2006.

[9] B. S. Guy and W. E. Patton, “The Marketing of Altruistic Causes: Understanding Why People Help,” The Journal of Consumer Marketing, vol. 6(1), 1989, pp. 19–30, doi:10.1108/EUM0000000002536.

[10] P. Hankinson, “Brand Orientation in Charity Organizations: Qualitative Research into Key Charity Sectors,” International Journal of Nonprofit and Voluntary Sector Marketing, vol. 5(3), 2000, pp. 207–219, doi:10.1002/nvsm.114.

[11] S. Hibbert and S. Horne, “Giving to Charity: Questioning the Donor Decision Process,” Journal of Consumer Marketing, vol. 13, 1996, pp. 4–13, doi: 10.1108/07363769610115366.

[12] R. Kaplinsky and M. Morris, A Handbook for Value Chain Research. Report prepared for IDRC, University of Sussex Institute of Development Studies, 2001.

[13] M. Kleijnen, K. De Ruyter, M. Wetzels, “An Assessment of Value Creation in Mobile Service Delivery and the Moderating Role of Time Consciousness,” Journal of Retailing, vol. 83(1), 2007, pp. 33–46, doi: 10.1016/j.jretai.2006.10.004.

[14] Y. Kuo, C. Wu and W. Deng, “The Relationships among Service Quality, Perceived Value, Customer Satisfaction, and Post-Purchase Intention in Mobile Value-Added Services,” Computers in Human Behavior, vol. 25(4), 2009, pp. 887–896, doi: 10.1016/j.chb.2009.03.003.

[15] S. Lee, “Marketing charities in the 1990s,” in The Henderson Top 1000 Charities: A Guide to UK Charities, R. Henderson, Eds. London, UK: Hemmington Scott, 1993.

[16] N. Mallat, “Exploring Consumer Adoption of Mobile Payments – A Qualitative Study,” Proc. Helsinki Mobility Roundtable, Helsinki, Finland, June 2006.

[17] B. Martin and P. Nightingale, “Introduction,” in The Political Economy of Science, Technology and Innovation, B. Martin and P. Nightingale, Eds. Cheltenham, UK: Edward Elgar, 2000, pp. 13–42.

[18] B. Nooteboom, “Service Value Chains and Effects of Scale,” Service Business, vol. 1(2), 2007, pp. 119–139, doi: 10.1007/s11628-006-0009-4.

[19] P. A. Pavlou, “Consumer Acceptance of Electronic Commerce. Integrating Trust and Risk with the Technology Acceptance Model,” International Journal of Electronic Commerce, vol. 7(3), 2003, pp. 101–134.

[20] M. Porter, Competitive Advantage: Creating and Sustaining Superior Performance. New York, NY: Free Press, 1985.

[21] K. Pousttchi and D. G. Wiedemann, What Influences Consumers’ Intention to Use Mobile Payments? New York, NY: Free Press, 2007.

[22] W. W. Royce, “Managing the Development of Large Software Systems: Concepts and Techniques,” Proc. 9th International Conference on Software Engineering (ICSE 87), IEEE Computer Society Press, 1987, pp. 328–338.

[23] J. Saxton, “A Strong Charity Brand Comes from Strong Beliefs and Values,” Journal of Brand Management, vol. 2(4), 1994, pp. 211–220.

[24] P. G. Schierz, O. Schilke and B. W. Wirtz, “Understanding Consumer Acceptance of Mobile Payment Services: An Empirical Analysis,” Electronic Commerce Research and Applications, vol. 9(3), 2010, pp. 209–216, doi:10.1016/j.elerap.2009.07.005.

[25] C. Sørensen, Enterprise Mobility: Tiny Technology with Global Impact on Work. Technology, Work and Globalization Series, Hampshire, UK: Palgrave Macmillan, 2011.

[26] UK Charity Commission, Charities Act 2006. [online] <http://www.legislation.gov.uk/ukpga/2006/50/contents> [accessed 29 February 2012].

 ISBN: 978-1-61208-229-5

A Privacy Preserving Range Extension for Commercial WLANs with User Incentives

Johannes Barnickel
IT Security Group
RWTH Aachen University
barnickel@umic.rwth-aachen.de

Ulrike Meyer
IT Security Group
RWTH Aachen University
meyer@umic.rwth-aachen.de

***Abstract*—Worldwide service availability via international roaming is one of the success factors of mobile telecommunications and hopefully also for WLAN access in the near future. Recently, a promising protocol suite for inter-operator roaming in commercial WLAN has been proposed. This protocol suite offers several advantages over other roaming protocols such as secure payment, short time tariff shaping, and strong privacy guarantees. In this paper, we propose an extension to this protocol suite, which allows any WLAN customer with a mobile device that supports virtual interfaces on its WLAN card to act as a paid relay station, or as we call them, Hops. The WLAN provider profits from these relaying stations at they increase the coverage area of his access points even beyond his domain. The owner of the Hop will receive monetary compensation over an integrated tick payment scheme. Like the original protocol suite, our protocol extension offers secure payment, tariff shaping, and strong privacy guarantees.**

keywords – WLAN roaming; micropayment; Internet access; mobile Internet; wireless hops.

I. INTRODUCTION

Worldwide roaming is one of the most valuable services provided by modern mobile operators. It is based on the fact that each mobile device (**MD**) has a contract with one of several home networks (**HN**s), which has MD's billing information. The HN has a roaming agreement with various foreign networks (**FN**s) via which they agree to provide access services to each other's customers. MDs are billed for the use of FN's service via HN, and roaming tariffs are negotiated between the two providers. Unfortunately, the latter has led to very high roaming prices and users ending up with unexpectedly high bills due to a missing transparency of the tariffs being charged. In addition, tariffs are quite inflexible and need to be fixed based on legal agreements between the operator rather than being based on the current demands. Finally, current roaming practices do not preserve the privacy of customers as HN learns everything about MD's service use via FN, and FN learns the identity of MD. Some mobile operators have started to use SIM-based access to WLANs and are thereby able to reuse the roaming infrastructure of their telephony networks in the WLAN context – including its shortcomings.

In many commercial WLANs, however, user are still directed to a webpage where they have to provide their credit card information to the operator of the access point, which is cumbersome for short term use and requires the user to disclose his personal data. Also, the credit card transaction fees make paying small amounts for Internet access not efficient. Often, long term contracts are offered by operators of multiple access points, e.g., mobile phone operators, or dedicated providers. For the user, this means having to sign an (often long running) contract with an unknown provider, without being able to judge how often he will be close to an access point of this provider.

The roaming solution proposed in [1] addresses these shortcomings of current roaming solutions. It combines secure and convenient access to paid WLANs with tariff transparency, tariff flexibility, integrated micropayment, and privacy protection. Unlike in mobile telephony networks, the FNs are able to change the tariffs they offer at any time without even notifying HN, as they are not part of the roaming agreements. As MDs can choose any tariff offered by any FN, tariff negotiation between MDs and FNs is enabled. To retain customer privacy, the HN does not receive any details about its clients' sessions, and the FN will not be able to identify or track HN's clients.

In this paper, we propose a new Hop extension to the roaming solution proposed in [1]. This proposed extension allows any MD connected to a participating WLAN to act as access point itself. We refer to such an MD as "Hop". These Hops increase the area covered by WLAN, increase the operator's number of potential clients, and can help to create ubiquitous access. The owner of a Hop is reimbursed for acting as a Hop such that our approach does not suffer from missing incentives to share connectivity like many free WLAN initiatives [2] do. A client using a Hop will still receive a single bill from his home network. Note that acting as a Hop is perfectly feasible with off the shelf laptops and smartphones. Many mobile devices today support virtual interfaces in their WLAN module or can use multiple different network interfaces at the same time. This means that these devices are able to act as client and access point in different networks at the same time.

The rest of the paper is structured as follows: In Section II, we reconsider the approach to WLAN roaming described in [1]. In Section III, we extend the protocol to cover mobile devices acting as Hops. We review related work in Section IV.

 ISBN: 978-1-61208-229-5

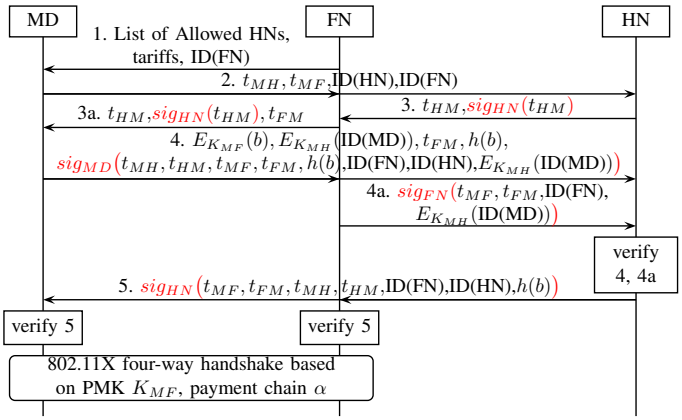


Fig. 1. Basic Connection Setup Protocol

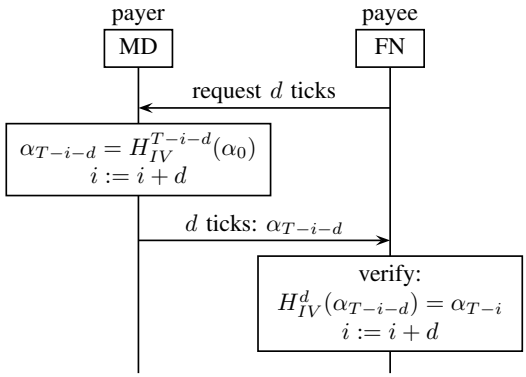


Fig. 2. Basic Tick Payment Protocol

II. The Basic Roaming Protocol

In this section, we briefly resume the roaming protocol suite proposed in [1]. We will then extend it to include Hops in Section III.

It is assumed that HN and MD, and HN and FN are in possession of each other's public key. The protocol suite consists of three protocols: a connection setup protocol, a tick payment protocol, and a clearing protocol. The connection setup protocol involves MD, HN, and FN and is used for authentication, key agreement, and payment initialization when MD requests access to an access point operated by FN. The tick payment protocol is used between MD and FN. It keeps the connection alive with regular tick payments created by MD to pay for the service it uses. When the connection ends, the clearing protocol is executed between FN and HN, and optionally MD.

A. Basic Connection Setup Protocol

The connection setup protocol assures FN that MD is owned by a client of HN, and that HN will pay FN for the services MD used. It allows FN to advertise his current tariffs and these tariffs are authenticated as part of the setup protocol such that MD is assured of the tariffs offered by FN and that FN is a roaming partner of HN. HN is assured of MD's identity and that MD has agreed on the tariff used. At the same time, MD's privacy is protected, such that HN does not learn payment details from the connection setup protocol. FN cannot learn MD's identity, cannot recognize if MD has been a client before, and therefore cannot track MD using access points in multiple locations. Obviously, MD would also have to change its MAC address for every new connection to a FN it used before.

The protocol is described in Fig. 1, and the notations are given in Table I.

Message 1 is a broadcast that is continuously sent by FN's access point. It contains a list of tariffs and HNs with which FN has a roaming agreement, i.e., whose clients may connect. Tariffs can be offered in cost per minute or cost per data volume. The broadcast mechanism allows FN to change its tariffs at any time and it allows MDs to discover the network and to review the tariffs offered by FN. We include the ID of FN in message 1 as well, although it was not mentioned explicitly in [1]. The client software on the MD is able to decode this broadcast, and enables the user to select one of the tariffs offered by FN. When multiple FNs offer wireless networks in the same location, the user can choose freely among the FNs.

We have implemented the broadcast on commodity hardware using a special encoding in the SSID. The SSID may contain up to 256 bits. The SSID must contain the FN's name, which could be at most 10 characters = 80 bits. With four different interval sizes and 32 different per unit prices (e.g., from 0.1 to 28.5ct in steps of 20%), 8 bits per tariff offered are required. With two types of tariffs (data volume and time based), up to 8 tariffs are possible for each network, requiring $8 \cdot 8 = 64$ bits of the SSID. Allowed HNs can be encoded as a bit string (assigned by some authority), the length of which could be 10 bit, allowing for a total of $2^{10} = 1024$ HNs. With an operator name length of 10 bytes in clear text, at most $\lfloor (256 - 80 - 64) : 10 \rfloor = 11$ allowed HNs fit in the SSID.

Messages 2.–5. contain a key establishment protocol based on Diffie-Hellman with authentication via signatures. As MD and FN do not share public keys, HN verifies the signatures and confirms their correctness in message 5. MD and FN establish the key K_{MF}, of which one derivation will be used in a subsequent 802.11X four-way handshake, and another derivation is used to hide details about the payment information from HN. MD and HN establish the key K_{MH} to hide MD's identity ID(MD) from FN. Message 3. and 3a. are modified over [1] to include HN's signature to prevent an attacker acting as FN from executing a man-in-the-middle-attack on K_{MH} to extract ID(MD).

Messages 4. and 5. also contain the list of offered tariffs, the tariff selected by MD, payment initialization values IV and α_T, and the first tick payment, which are summarized as b in Fig. I. Further service intervals and clearing of payments are discussed in Sections II-B and II-C.

B. Tick Payment Protocol

The tick payment protocol has been proposed by Horn and Preneel [3] and Pedersen [4]. In [1], the first tick payment has been integrated into the setup phase to speed up service provision.

 ISBN: 978-1-61208-229-5

Diffie-Hellman related:	
g	Publicly known generating element of a finite group G where the discrete logarithm problem is hard
p	Publicly known large prime
r_{XY}	Private DH key chosen by party X for key setup with party Y
t_{XY}	Public DH key calculated by X for setup with Y: $g^{r_{xy}} \bmod p = t_{xy}$
General Cryptographic Operations:	
$H_{IV}(m)$	Preimage resistant hash function with input m and initialization vector IV
K_{XY}	Symmetric key established between parties X and Y during the protocol run
$sig_X(m)$	Signature of m by party X, does not include the message m. Signatures are unlinkable with regard to the signer.
$E_K(m)$	Symmetric encryption of plaintext m with key K, e.g., AES
ID(FN)	Identifier of FN, i.e., its unique brand name
Payment Related:	
tariffs	List of: type of tariff (per data volume, time, packets, etc), price (amount, currency, unit), total ticks T (connection limit), ticks per unit d, e.g., charged per time, 0.01 EUR per 30 seconds, 14400 ticks total, 5 ticks per unit
d	Amount of ticks per unit as requested per tariff
α	payment chain used between MD and FN
α_0	Root of the payment hash chain chosen by the payer
α_T	Last element in the chain in generation order, $\alpha_T = H_{IV}^T(\alpha_0)$
α_{T-d}	First tick payment. $\alpha_{T-d} = H_{IV}^{T-d}(\alpha_0)$
IV	Initialization vector chosen by the payer
b	Payment info vector. $b = (IV, \alpha_T, \alpha_{T-d},$ selected tariff, offered tariffs)
pay_α	graceful payment string $pay_\alpha = \alpha_T, IV, T, sum, \alpha_{end}$
Hop Related (Section III):	
htariffs	tariffs offered by FN which are supported by Hop
PID(Hop)	pseudonym ID of Hop, i.e., a permanently fixed random string shared with HN
α	payment chain used between Hop and FN
β	payment chain used between MD and FN
γ	payment chain used between MD and Hop
c	Payment info vector. $c = (IV_\beta, \beta_T, \beta_{T-d}, IV_\gamma, \gamma_T, \gamma_{T-d},$ selected tariff, offered tariffs)
t_*	All public DH keys $t_* = (t_{GM}, t_{MG}, t_{FM}, t_{MF}, t_{HopM}, t_{MHop})$

TABLE I
NOTATIONS

Initialization of Tick payment: MD generates payment data by randomly choosing α_0, IV, and then calculating a payment chain α, where $\alpha_i = H_{IV}(\alpha_{i-1}), i \in \{1, \ldots, T\}$, where T is given by the tariff, and IV and α_0 are randomly chosen by the MD. MD commits to the payment by calculating a signature on α_T, IV, the ID(FN), and the selected tariff in message 4 of the setup protocol, so that FN can later prove to HN that it was MD who created the payment for FN. FN verifies the first tick α_{T-d} by testing $H_{IV}^d(\alpha_{T-d}) = \alpha_T$. If successful, FN provides service to MD until the first service interval is used.

For **Later Service Intervals**, e.g., after i ticks, FN will request d new tick payments as illustrated in Figure 2. After MD has sent the last tick α_{T-i-d}, FN verifies that $H_{IV}^d(\alpha_{T-i-d}) = \alpha_{T-i}$. Both parties increase i by d and store i. This can be repeated until $i > T$. After $\frac{T}{d}$ service intervals, all ticks have been used and the connection aborts. Therefore, T limits the amount of service used in a session. Note that T is chosen by FN.

d ticks correspond to a small amount of money called unit, which should be chosen so small that losing it is not a problem, because at most one unit will be lost when the connection aborts unexpectedly, or when FN provides no service, e.g., d ticks could be worth 0.05 EUR or less. We use $d = 1$ in our implementation.

C. Basic Clearing Protocol

There are two separate clearing protocols that differ in the information that is kept private from HN.

The **abort protocol**, as shown in Figure 3 is started by FN when MD does not actively terminate the connection, or in general when MD fails to initiate the graceful ending protocol (discussed later). FN has obtained the signature from MD in message 5 during the connection setup protocol described in Section II-A, and the last tick payment α_{T-i} during the tick payment phase described above. By sending the setup messages 3, 4, 5, b, and α_{T-i}, FN can prove to HN that MD is a customer of HN, the amount of provided service, and the selected tariff, which results in the amount to be paid. HN will reimburse FN and charge MD. Regarding privacy, HN will obtain knowledge of the tariffs offered by FN, the tariff MD and FN have agreed on, and the amount of service MD used.

The **graceful ending protocol**, as shown in Figure 4 is started by MD when it does not want to use further service. At the end of a connection with an amount of service used worth sum, MD creates a payment string for FN,

 ISBN: 978-1-61208-229-5

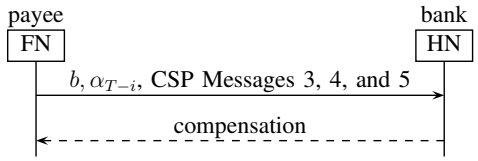


Fig. 3. Basic Clearing Phase after Abort. CSP = Connection Setup Protocol

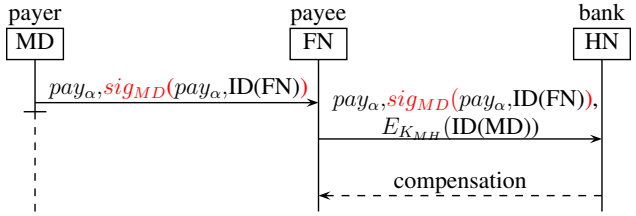


Fig. 4. Basic Clearing Phase with Graceful End

$$pay_\alpha = \alpha_T, IV, T, sum, \alpha_{end},$$

which is used in the ending message pay_α, $sig_{MD}(pay_\alpha$, ID(FN)) for FN. FN will forward this signature to HN. The information contained in this message is sufficient for FN to prove to HN that MD owes FN the amount in question. In the interest of MD's privacy, HN will not obtain knowledge of any tariff or service use details MD and FN have agreed on. As FN cannot verify the signature on pay_α, ID(FN), it keeps the data required to run the abort protocol, so that it can be executed when HN refuses the graceful ending message. E.g., when MD tries to cheat by sending an invalid signature, FN will behave as if MD aborted the connection. Note that given α_i, no one can calculate α_{i-j} for any $j > 0$ because H_{IV} is a preimage resistant hash function. Therefore, new ticks to an existing chain cannot be forged. Due to the nature of tick payment, MD only loses the value of a single (small) service unit when FN stops providing service after MD has paid for it. To avoid HN reconstructing service use from the amount of ticks and the total sum, FN can vary his service interval size, and adjust the price per unit so that the effective price stays constant.

III. Extension to Hops

In this section, we describe our new Hop extension of the protocol suite described in the last section. It allows MDs to act as Hops, i.e., access points for other MDs. Hops will receive a small payment from MD for providing service, and FN will receive the regular payment as if MD was connected without a Hop. Note that the extension is not straight forward, as Hops — as opposed to FNs — do not have a trust relationship with the HNs of other MDs. Also, there is a higher risk that Hops act maliciously than that FNs act maliciously as the later can be considered to care about their reputation as they want to stay in business.

A. Scenario and Requirements

In the following, the term MD is always used for a device that uses a Hop and has no direct wireless connection to FN. The Hop is a regular MD currently connected to FN, and is owned and operated by an end user. MD's home network will be called **GN** (guest network) in the following, and the home network of Hop will be called **HN**. No roaming agreement is required between HN and GN, but both need a roaming agreement with FN.

As a Hop is using resources to provide service to a MD (battery life, system load), an incentive is required for MDs to become Hops. This is achieved by the MD paying a small fee to the Hop with each tick payment. MD also pays the regular fees for the services it uses to FN.

The amount paid by MD to Hop is chosen by FN and advertised in FN's tariff broadcast. Not allowing the Hops to freely choose the amount they earn prevents them from charging disproportionate tariffs from careless MDs, which may ruin FN's reputation as well. Note that a Hop can choose to accept or reject individual tariff options advertised by FN before forwarding them to MD. Depending on the tariffs offered by FN, acting as a Hop might even be a business model, i.e., other providers might deploy fixed devices to act as Hops in highly frequented places along the borders of FN's network coverage.

MDs have to pay more for service used over a Hop compared to a direct connection. If the incentive for Hop would be paid by FN, it would be subtracted from FN's profit, and enable attackers running an MD to pose as both a Hop and an MD at the same time, e.g., by using two devices, and pocketing the Hop incentive themselves. Theoretically, FN could also setup Hops and try to charge MDs more, however, we neither consider this realistic nor an attack per se, as FN can freely set its tariffs anyway.

Our extended roaming protocol suite aims at meeting the following goals:

Sec-1: Authentication and key establishment between MD, Hop, FN, and GN.
Sec-2: MD can avoid a Hop that acted dissatisfactory in the past.
Sec-3: Hop cannot read or modify MD's traffic.
Pri-1: MD must stay anonymous and untrackable to anybody.
Pri-2: Hop must stay anonymous to anybody.
Pri-3: GN must not learn details about MD's and Hop's session with FN.
Pri-4: HN must not learn details about MD's and Hop's session with FN.
Pay-1: Hop will never have to pay for the services MD uses with FN.
Pay-2: FN and Hop will be paid by MD for the services MD uses.
Pay-3: FN and Hop cannot charge more than negotiated with MD.

B. Extended Connection Setup Protocol

The Hop Protocol is illustrated in Figure 5 using the notations from Table I.

 ISBN: 978-1-61208-229-5

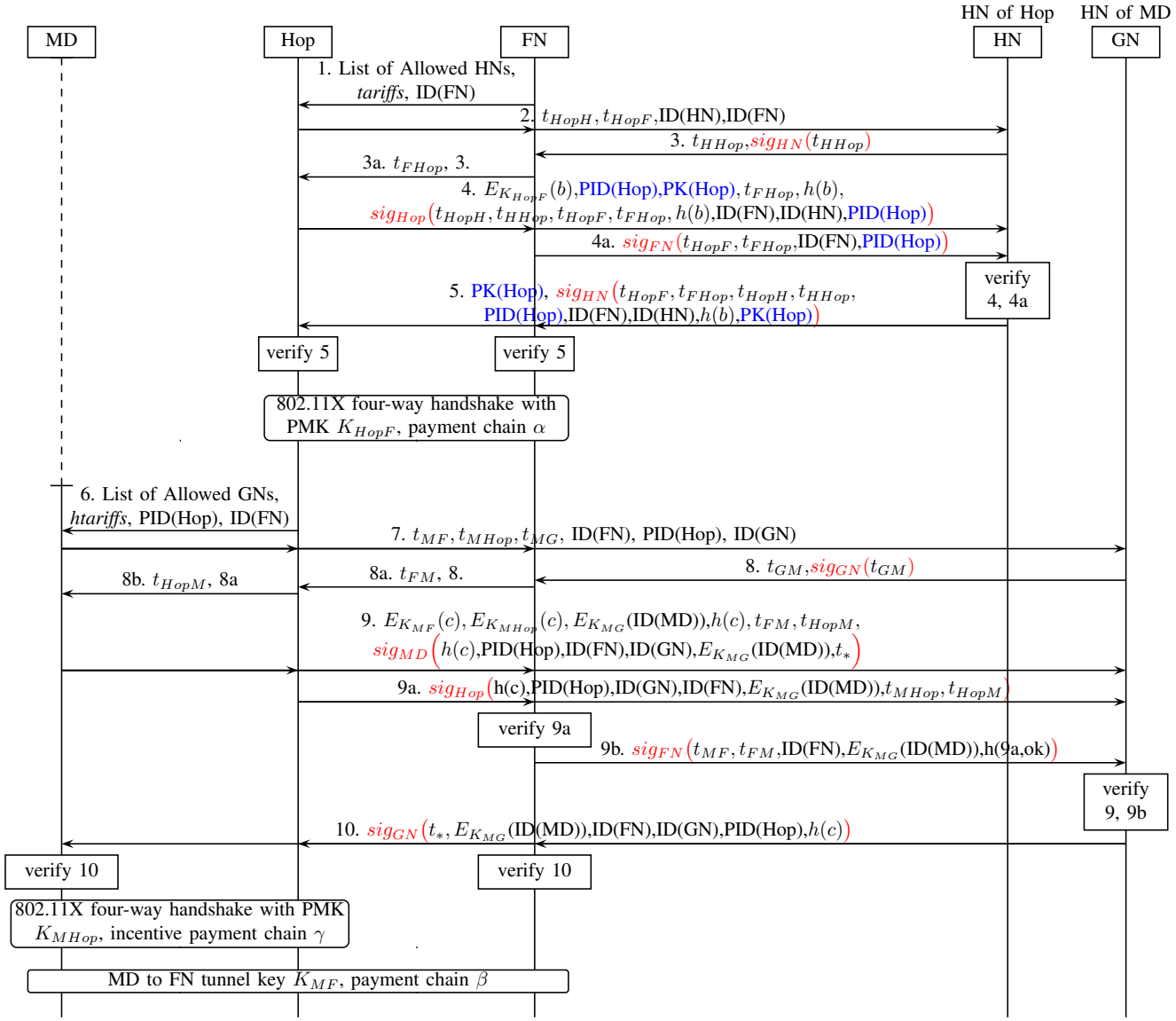


Fig. 5. Extended Connection Setup Protocol. red: signatures; blue: changes to 1 in 1.–5.

1.–5. These messages are similar to the Basic Connection Setup Protocol as discussed in Section II-A, except for the added pseudonym ID of Hop PID(Hop), and its public key PK(Hop). The tariff list now also includes rates indicating how much Hops must be paid. HN confirms that Hop is entitled to act as a Hop by confirming its PID(Hop).

5.–11. These messages contain two entwined runs of the roaming mechanism discussed in Section II-A between the pairs MD, Hop and MD, FN. These are similar to the two-party protocol, but now the Hop uses the static identifier PID(Hop) issued by HN instead of the encrypted identifier $E_{K_{MH}}$(ID(MD)) that MD used in the basic protocol. This enables MD to recognize and avoid certain Hops. Two new tick payment chains β between MD and FN, and γ between MD and Hop are initialized. The mechanism is straightforward and works as in the two-party case. Payment information during the setup protocol is shortened to c in Figure 5.

6. Like FN, the Hop continuously sends a broadcast message, e.g., using a special encoding in the SSID. It contains the pseudonym ID of Hop PID(Hop) issued by HN of Hop, a (possibly) reduced set of tariffs from FN's broadcast called *htariff*, the ID of FN, and a list of allowed GNs (HNs of MDs), which are taken from FN's broadcast (1.).

7. The user of MD selects a suitable tariff. MD chooses DH private values $r_{MHop}, r_{MF}, r_{MG} \in_R \mathbb{Z}_p$, and calculates public $t_{MHop} = g^{r_{MHop}}, t_{MF} = g^{r_{MF}}, t_{MG} = g^{r_{MG}} \mod p$. t_{MHop} is meant for key establishment with Hop, t_{MF} for FN, and t_{MG} for GN. MD sends the IDs of Hop and FN, the public DH values to Hop, and the identifier ID(GN) of its HN so that FN will know where to forward messages 7 to. Hop forwards the message to FN when PID(Hop) is correct. FN verifies that it has a roaming agreement with GN, and forwards the message if it does.

8. GN verifies that it has a valid roaming agreement with FN. GN creates private $r_{GM} \in_R \mathbb{Z}_p$ and public $t_{GM} = g^{r_{GM}} \mod p$ and calculates $K_{MG} = {t_{MG}}^{r_{GM}} \mod p$ for use with MD. t_{GM} is sent to FN.

8a. FN creates private $r_{FM} \in_R \mathbb{Z}_p$ and public $t_{FM} = g^{r_{FM}} \mod p$ and calculates $K_{MF} = {t_{MF}}^{r_{FM}} \mod p$ for use with MD. t_{GM}, t_{FM} are sent to Hop.

8b. Hop creates private $r_{HopM} \in_R \mathbb{Z}_p$ and public $t_{HopM} = g^{r_{HopM}} \mod p$ and calculates $K_{MHop} = {t_{MHop}}^{r_{HopM}} \mod p$ for use with MD. t_{GM}, t_{FM}, t_{HopM} are sent to MD.

9. MD calculates K_{MF}, K_{MHop}, K_{MG}. MD generates the payment data c according to the tariff selected, which contains the payment chains β for FN and γ as incentive for Hop. c also contains the tariffs offered by Hop and an identifier of the tariff selected by MD. MD creates a signature on its encrypted identifier, a hash of c, all the ephemeral DH public parameters t_*, the ID of FN and GN, and PID(Hop). MD sends this signature, its identifier encrypted for GN, the payment data c encrypted for Hop and FN, a hash of c, and the two ephemeral DH public parameters not seen by HN so far t_{FM}, t_{HopM} to GN. This data is required by GN to verify the signature.

9a. Hop verifies $h(c)$. Hop creates a signature on MD's encrypted identifier, the ephemeral DH public parameters Hop used t_{MHop}, t_{HopM}, the ID of FN and GN, and PID(Hop). Hop sends this signature and message 9 to FN.

9b. FN verifies $h(c)$ and the signature of Hop from message 9a using the public key from message 4, which was confirmed by HN in message 5. FN creates a signature on MD's encrypted identifier, the ephemeral DH public parameters FN used t_{MF}, t_{FM}, the ID of FN and GN, and PID(Hop). FN sends this signature, message 9, and 9a to GN.

10. After GN verifies the signature by MD and FN from message 9 and 9b, GN creates a signature on all ephemeral DH values, the identifiers of GN, FN, Hop, and the encrypted identifier of MD, and the hashed payment information $h(c)$, which is sent to FN. FN verifies the signature by GN and forwards it to Hop when the verification succeeds. Hop forwards the message to MD, who verifies GN's signature.

Now that the parties have authenticated, established keys and initialized payment, MD and Hop execute an 802.11X handshake using a derivation of K_{MHop} and the payment chain γ, and MD and FN set up an IPsec tunnel using a derivation of K_{MF} and the payment chain β.

C. Discussion of the Extended Connection Setup Protocol

The Hop connection setup protocol is built on similar goals as to the basic roaming mechanism described in [1], on which we gave a summary in Section II-A. We will now discuss how the security and privacy goals for the extended protocol described in Section III-A are achieved.

Sec-1: Secure authentication and key establishment between MD, Hop, FN, and GN is achieved as all parties include ephemeral public keys from messages 2–3, 7–8 within the signed parts of messages 4–5, 9–10. The signature that cannot be verified directly due to lacking public keys are verified by parties that are trusted by MD (GN verifies FN's signature), Hop (FN is trusted via HN's roaming agreement, FN verifies GN's signature), and GN (FN verifies Hop's signature). FN verifies Hop's signature using the PK(Hop) supplied in message 4 and confirmed by HN in message 5. Therefore, all parties are aware that the other parties are actively participating in the current protocol run. The keys K_{MHop}, K_{MF}, and K_{MG} established during the protocol run are fresh, as the ephemeral public DH parameters are chosen by all parties for only this session. Also, the keys are **exclusive** as they can only be calculated by a party that knows the corresponding private ephemeral DH parameter r corresponding to the public parameter t it sent. Explicit key confirmation is achieved by the encryption of c between MD and Hop, and MD and FN, and by the encryption of ID(MD) between MD and GN. Thus, mutual belief in the keys K_{MHop}, K_{MF}, and K_{MG} is achieved. Note that there is always input from at least one self chosen ephemeral DH value in every signature in the protocol to prevent **reuse of old signatures** by an attacker.

Sec-2: PID(Hop) is sent to MD to achieve **linkability** of Hop to MD. This way, MD is able to avoid using Hop when service has been poor before.

Sec-3: The Hop cannot read or modify traffic between MD and the Internet, because the traffic between MD and FN is encrypted and integrity protected using an IPsec tunnel based on a derivation of the key K_{MF}.

Pri-1: The MD stays anonymous and untrackable to both Hop and FN, as ID(MD) is only sent encrypted with K_{MG}, which is only known to GN.

Pri-2: PID(Hop) is issued by Hop's HN and does not contain a real name, so that Hop stays anonymous, but linkable.

The **verification of MD's signature** on $h(c)$ sent in message 9 is interesting. Only GN is able to verify MD's signature directly. GN signs the $h(c)$ sent by MD in message 10, which can be verified by FN, but not by Hop. Therefore, another mechanism is needed. Hop includes $h(c)$ in its signature in message 9a, which FN verifies. After verifying message 9a and 10, FN knows that MD has encrypted the same c for Hop, FN, and GN.

D. Payment for Hops

The tick payment protocol uses two payment chains β from MD to FN and γ from MD to Hop. The payment chains are bound to the authenticated payer and the intended receiver by MD's signature in message 9 of the Extended Connection Setup Protocol (Figure 5). As shown in Figure 6, FN requests d new ticks after a service interval has been used up by MD. MD is paying to FN and Hop by sending ticks β_i and γ_i.

MD keeps track of its service use so that it cannot be overcharged by FN. Hop keeps track of MD's service use and verifies that payment ticks γ_i arrive in a timely fashion. Hop

 ISBN: 978-1-61208-229-5

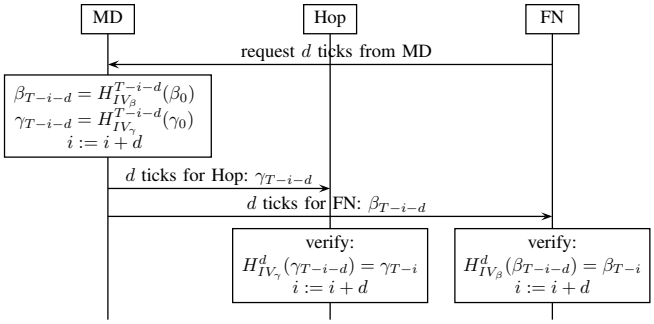


Fig. 6. Extended Tick Payment Protocol

and FN verify the received ticks in the same way using H_{IV}. The Hop does not have to request payment itself, as the tariff chosen by MD includes the incentive paid to Hop, which uses the same unit size (time or data volume) and maximum number of service intervals T.

E. Extended Clearing Protocols

FN is clearing the payments α from Hop to FN as described in Section II-C. Only the payments from MD to FN β and from MD to Hop γ are cleared using the extended clearing protocols. As for the basic protocol, there are two variants, depending on whether or not MD sent a graceful ending message.

The Extended Abort Protocol is executed when MD aborts the connection as shown in Figure 7. The messages from the extended connection setup protocol are used to prove to GN that MD has committed on the payment chains to pay FN and Hop, as they contain MD's signature on c and GN's signature of $h(c)$. Hop and FN disclose c to GN, who can verify it using $h(c)$. The last tick payments β_{end} and γ_{end} and the tariff data from c allow FN and Hop to prove the amount to be paid.

By disclosing c and the last tick payment, the GN will obtain knowledge of the tariffs offered by FN, the tariff MD has agreed on, and the amount of service used by MD, which can be avoided by MD sending an ending message.

As the Hop can lose its wireless link at any time, e.g., when the user forgets to log off and leaves the range of FN's wireless network, the abort clearing protocol can be executed in regular intervals with a delay flag so that FN will not contact GN immediately.

The Extended Graceful Ending Protocol is shown in Figure 8. When MD does not want to use further service, it sends an ending message for Hop, which is relayed to FN. The message is based on two signed payment strings,

$$pay_\beta = \beta_T, IV_\beta, sum_\beta,$$
$$pay_\gamma = \gamma_T, IV_\gamma, sum_\gamma,$$

to pay for services MD used itself. $\beta_T, \gamma_T, IV_\beta, IV_\gamma$ are random values that the payment chain is based on. They are sent to prevent double spending so that Hop and FN cannot clear the same payment chain twice. sum_β and sum_γ is the amount to be paid to FN and Hop in a real world currency. Hop forwards message 1, but also the last tick payment Hop received γ_{end} to FN. FN is forwarding message 1 to GN along with MD's encrypted identifier used in the extended connection setup protocol. GN verifies the signature of MD and acknowledges the claim. GN will credit FN, possibly later at the end of a billing period. GN cannot credit Hop, because GN might not have a roaming agreement with HN. Therefore, GN sends payment for Hop to FN, and in message 4 FN forwards it to HN, who credits Hop in message 5.

The Hop has included γ_{end} in message 1a. to FN so that FN can execute the Abort Clearing Protocol without contacting the Hop again, should GN reject MD's signature from message 1. The other information c, β_{end} and the Setup Message 7, 8, 9, 10 are already known to FN from the Extended Setup Protocol. Hop includes c again to identify the connection with MD.

F. Discussion of the Extended Payment Protocol

The tick payment chains for MD to Hop and MD to FN payments are both securely initialized in the Extended Connection Setup Protocol. Each of the payment chains provides the properties discussed in Section II-B such that the chains cannot be forged, payments cannot be stolen and cleared by someone else, and payments cannot be used or cashed more than once. We will now discuss how the payment security and privacy goals described in Section III-A are achieved.

Pri-3: GN does not learn any details about MD's and Hop's session with Hop and FN when the extended graceful end protocol is executed correctly, as the payment strings pay_β and pay_γ only contain the amount to be paid and the party to be paid. However, if GN would wrongfully reject MD's signature, it can force FN to reveal the details. FN would detect this attack if it happens often and could cancel the roaming agreement with GN.

Pri-4: HN does not learn any details about MD's and Hop's session with Hop and FN, because HN only receives ID(GN) and the payment strings, which only contain the sum to be paid, Hop's PID, and ID(GN).

Pay-1: The Hop is assured that it will not have to pay for the services MD uses with FN, because MD is using its own payment chain β with MD, and FN counts the service used by MD separately from those used by Hop. When FN tries to overcharge Hop, Hop can abort the connection upon receiving a wrongful tick payment request. The maximum risk for Hop is the value of a single tick payment.

Pay-2: FN and Hop are convinced that they will be paid by MD for the services MD uses, because MD has committed on one payment chain each for Hop and HN in message 9, which was confirmed by GN in message 10, which was confirmed to Hop by FN forwarding message 10 and providing subsequent service to MD. Every single tick payment sent by MD can be verified by Hop and FN immediately, and clearing does not rely on MD's cooperation.

Pay-3: FN and Hop cannot charge more than negotiated with MD. They cannot calculate additional tick payments in pay_β and pay_γ, because H_{IV} is a one-way function. FN and Hop cannot clear the same connection twice, as β_T, IV_β and/or γ_T, IV_γ will be the same as those cleared before, which will be detected and rejected by GN.

 ISBN: 978-1-61208-229-5

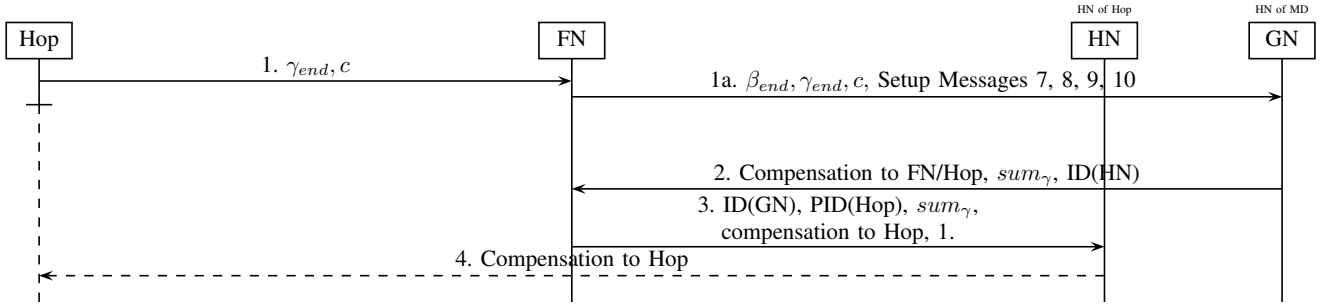


Fig. 7. Extended Abort Clearing Protocol

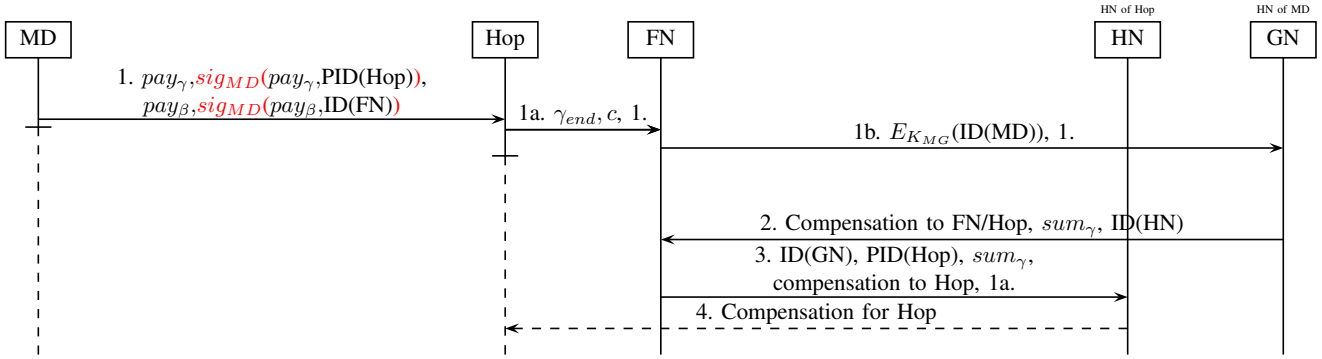


Fig. 8. Extended Clearing with Graceful End

IV. Related Work

In this section, we will compare the proposed extended solution to existing academic and non-academic roaming approaches and show that none of these approaches simultaneously meets all the features our solution offers.

A. Roaming Solutions without Hops

3GPP [5] relies on stored customer profiles to facilitate billing and user authentication. Tariff selection on a per connection basis for users is possible with some operators by dialing special codes on the MDs, but no on-demand tariff shaping for operators. Despite the TMSI mechanism, active attackers are able to track mobile devices, HN always obtains all connection details, and FN always obtains the subscriber's longterm identifier.

A variety of roaming protocols exist that do not support payment initialization and tariff negotiation. These include for example the protocol suggested in [6], [7], [8], [9], [10], [11], [12], [13], existing solutions like the Extensible Authentication Protocol [14] in 802.11i WLANs, and the recently launched PassPoint by the Wi-Fi Alliance and Wireless Broadband Alliance [15]. We will only discuss protocols in more detail that include secure payment as well.

In Buttyán-Hubaux [16], a customer care agency provides tickets to mobile devices. These tickets can be used by the mobile device to roam to different networks. The protocol is preserving the privacy of the user to the visited network, but not to the customer care agency. There is a single tariff chosen freely by the involved stations at each new connection, but no influence from the user on the selected tariff.

EAP-TLS-KS [17] uses a key splitting unique for each FN, distributed decryption, and distributed signatures for mutual authentication of MD and FN, which trades network round trips for additional cryptographic operations. EAP-TLS-KS can include any accounting method based on the Buttyán-Hubaux-Protocol.

B. Roaming Solutions with Hops

The solutions discussed so far did not discuss connections established over other parties (Hops). This research area is generally covered by wireless mesh networks (WMNs), where independent stations are also routers, even when they have no other network interface. Our work is not part of a WMN architecture, as it is limited to a single Hop.

ARSA [18] is a roaming solution based on identity based cryptography, which is not widely available for implementation. Brokers, connected to each other and to the operators, are used so that no agreements between operators are needed. User aliases are used to achieve unlinkability to the operator. A micropayment scheme is included. The Hops are not paid by the mobile device, but by the FN, which is thought to be more efficient for a large number of Hops and computationally weaker mobile devices as it is placing more load on the FN. Tariffs are announced, but only a single tariff priced per data volume is available per operator. Our approach avoids brokers, as all the participants would have to settle on the same one and would have to pay them a share, and rather uses the HN with a connection to the FN and a bilateral agreement, which are easier to set up.

The solution by Pierce-O'Mahony [19] combines roaming in GSM multi-mop networks with multiparty micropayment. Two MDs are connected to each other over a number of hops, and the initiator pays a large amount to the first hop, which keeps some of it, and forwards the rest to the next

 ISBN: 978-1-61208-229-5

hops, who repeat this process. The system is prepaid. MD's demand regarding QoS influences the tariff, but MD cannot directly choose a tariff, as the tariffs are chosen by the hops. The system is single-operator, which is hard to establish for a large audience in the real world. There is no protection against tracking of MD.

Jakobsson et al. [20] encourage collaboration in multi-hop networks using probabilistic micropayment, where the operator is capable of detecting and punishing misbehaving stations. Hops between MD and base station are paid for a random fraction of the packets they forward. The solution does not address tariffs, authentication, and privacy.

V. Conclusion and Future Work

We have presented an extension to the protocol suite in [1] for secure and privacy preserving roaming and payment in WLAN to include regular user devices acting as Hops, i.e., as relay stations to enhance the area where service is available. The privacy and security goals of the basic protocol suite are retained except for tracking of the Hop, which is a design choice to enable MDs to avoid certain Hops. The proposed solution retains tariff flexibility for users, Hops, and operators, as users can select a tariff that fits their demands. There is an incentive for Hops to provide service to MDs, and Hops only have to support tariffs they deem worthy. Operators are still free to modify their offered tariffs at any time. The clearing protocols ensure that all stations can be billed and credited correctly even when they disappear without advance notice or when they try to cheat.

We currently create a new EAP method for *hostapd* access points on Laptops, and a client for Linux and Android smartphones to implement the original protocol, and aim to implement the extension described in this paper as well. The client will be user friendly and recommend tariffs based on different Internet usage profiles, e.g., e-mail, chatting, surfing, and video chat.

We hope that our solution creates better WLAN coverage, fosters competition between paid WLANs operators, and ends insecure and cumbersome setup procedures.

VI. Acknowledgments

This work has been supported by the UMIC Research Centre, RWTH Aachen University. We want to thank the reviewers of Information Security Conference 2012 for providing insightful comments and discovering an attack on MD's anonymity in the original protocol.

References

[1] J. Barnickel and U. Meyer, "Security and privacy for wlan roaming with per-connection tariff negotiation," IEEE Conference on Local Computer Networks, 2011, pp. 338–353.

[2] WeFi, http://www.wefi.com/, retrieved August 1st, 2012, archived at http://www.webcitation.org/69Q5bO0zg.

[3] G. Horn and B. Preneel, "Authentication and payment in future mobile systems," Journal of Computer Security 8(2/3), pp. 183–207, 1999.

[4] T. Pedersen, "Electronic Payments of Small Amounts," Security Protocols, LNCS 1361, pp. 59–68, 1997.

[5] *3GPP TS 32.240 (Release 9): Telecommunication management; Charging management; Charging architecture and principles*, 3GPP Std., 2009.

[6] P. Bahl, S. Venkatachary, and A. Balachandran, "Secure wireless internet access in public places," IEEE International Conference on Communications, 2001.

[7] L. Buttyán, L. Dóra, F. Martinelli, and M. Petrocchi, "Fast certificate-based authentication scheme in multi-operator maintained wireless mesh networks," Journal of Computer Communications, Volume 33 Issue 8, pp. 907-922, May 2010.

[8] K. Bayarou, M. Enzmann, E. Giessler, M. Haisch, B. Hunter, M. Ilyas, S. Rohr, and M. Schneider, "Towards certificate-based authentication for future mobile communications," Wireless Personal Communications 29, pp. 283–301, 2004.

[9] J. Gu, S. Park, O. Song, J. Lee, J. Nah, and S. Sohn, "Mobile PKI: A PKI-Based Authentication Framework for the Next Generation Mobile Communications," Proceedings of ACISP'03, volume 2727 of LNCS, 180–191, 2003.

[10] T. Heer, S. Li, and K. Wehrle, "PISA: P2P Wi-Fi Internet Sharing Architecture," Seventh IEEE International Conference on Peer-to-Peer Computing, P2P2007, pp. 251–252, 2007.

[11] D. Leroy, G. Detal, J. Cathalo, M. Manulis, F. Koeune, and O. Bonaventure, "SWISH: Secure WiFi sharing," Computer Networks, Volume 55, Issue 7, 16 May 2011, pp. 1614–1630.

[12] M. Long, C.-H. Wu, and J. D. Irwin, "Localized authentication for wireless lan internetwork roaming," IEEE Conference on Wireless Communications and Networking, WCNC, pp. 496–500, 2004.

[13] L. Salgarelli, M. Buddhikot, J. Garay, S. Patel, and S. Miller, "Efficient authentication and key distribution in wireless IP networks," IEEE Wireless Communications Magazine Volume 10, Issue 6, pp. 52–61, December 2003.

[14] B. Aboda, D. Simon, and P. Eronen, "Extensible Authentication Protocol (EAP) Key Management Framework," IETF RFC 5247, 2008.

[15] WiFi Alliance, "Frequently Asked Questions on Wi-Fi CERTIFIED Passpoint," http://www.wi-fi.org/sites/default/files/uploads/20120626_Passpoint_FAQ.pdf, retrieved August] 2nd, 2012, archived at http://www.webcitation.org/69IOmvHOm.

[16] L. Buttyán and J. Hubaux, "Accountable anonymous Service Usage in mobile communication systems," EPFL SSC Technical Report No. SSC/1999/016, 1999.

[17] U. Meyer, J. Cordasco, and S. Wetzel, "An approach to enhance inter-provider roaming through secret sharing and its application to WLANs," Proceedings of the 3rd ACM international workshop on Wireless mobile applications and services on WLAN hotspots (WMASH), pp. 1–13, pp. 1–13, 2005.

[18] Y. Zhang and Y. Fang, "A secure authentication and billing architecture for wireless mesh networks," Wireless Networks, Volume 13, Number 5, pp. 663–678, 2007.

[19] M. Pierce and D. O'Mahony, "Flexible real-time payment methods for mobile communications," IEEE Personal Communications, Volume 6 , Issue 6, pp. 44–55, 1999.

[20] M. Jakobsson, J. Hubaux, and L. Buttyán, "A Micro-Payment Scheme Encouraging Collaboration in Multi-hop Cellular Networks," Financial Cryptography, 7th International Conference, FC 2003, pp. 15–33, 2003.

 ISBN: 978-1-61208-229-5

Sensing Learner Access to the Knowledge Spatially Embedded in the World

Masaya Okada
Graduate School of Science and Technology,
Shizuoka University
3-5-1, Jyohoku, Naka-ku, Hamamatsu, Shizuoka, Japan
email: m.okada@acm.org

Masahiro Tada
ATR Intelligent Robotics and Communication
Laboratories
2-2-2, Hikaridai, Keihanna Science City, Kyoto, Japan
email: mtada@atr.jp

***Abstract*—Real-world learning is an important application domain of mobile computing technologies. Real-world learning offers valuable opportunities for encouraging learners to acquire knowledge through experience in the world. Formative assessment by constant monitoring of intellectual achievement is an effective means of providing learners with adaptive support according to their situation. However, it is difficult to measure behavior and knowledge acquired in the real world, and no methodology has yet been developed to allow a formative assessment of learners in a real-world learning field (e.g., the natural environment). In this paper, we demonstrate that knowledge is three-dimensionally embedded in the world, and show a method for estimating how learners access such real-world knowledge. Our technology recognizes characteristic stay behavior and the associated body posture of learners, three-dimensionally estimates the target of their interest, and identifies the learning situation at any given time. As main achievements of our data analyses, we found that real-world knowledge is not only region dependent but also height dependent. We also showed that the learning topic of interest can be identified with wearable sensors (e.g., positioning sensors, 3-axis accelerometers, 3-axis gyroscopes, a barometer). These results are fundamental for realizing adaptive learning support based on systematic formative assessment.**

Keywords–mobile application; mobile learning; behavior recognition; context-aware service; real-world knowledge

I. INTRODUCTION

The ubiquity of computer technology is changing our daily activities and creating a next-generation society. Human intellectual activities in such a changing society are becoming more diverse, and more real-world oriented. To develop engineering technologies to enhance human activities, it is of fundamental importance to understand the dynamics of human knowledge processing. This paper focuses on people who interact with and learn in the real world, and makes a proposal for understanding how they access the knowledge that exists in the world.

A. Learning in and from the world

A cognitive architecture has traditionally been considered to be a process for forming a structured internal representation of the cognitive behavior of a person. In contrast, the theory of situated cognition postulates that a learner's cognitive process is embedded in the world, and proposes a cognitive architecture for acquiring knowledge through human interaction with the world [1], [2]. Real-world knowledge is embedded in situations [1], [2], and learners need to interact with the world, to have diverse experiences, and to autonomously investigate, find, and acquire real-world knowledge. Discovery learning through bottom-up knowledge acquisition is an important requirement of real-world learning. Environmental learning in nature is a typical example of real-world learning, and is selected as our model case.

B. Formative assessment of real-world learning

Formative assessment by constant monitoring of changes in learning situations is an essential factor for realizing adaptive learning support. Periodically testing the achievements of learners is a conventional method of formative assessment that is generally used in the case of desktop learning in a classroom. Learning management systems (LMS) assess learners' understanding by analyzing access logs of the LMS, and reviewing the results of tests and reports [3].

The main difference between desktop learning and real-world learning is that the effects of real-world learning depend on the outdoor experience of the learner. However, it is difficult to measure behavior and knowledge acquired in the real world, and no conventional research on real-world learning has proposed a methodology for carrying out a real-time evaluation of a learning situation that changes from moment to moment. There are no techniques to assess what real-world knowledge has been adequately acquired, and whether real-world learners have carried out meaningful activities for understanding their surroundings. These are problems that limit the effect of real-world learning.

The following are goals of the formative assessment method that we consider in the present research: (1) assessing how learners expand the areas that they are interested in, (2) assessing the time-series change in the degree that learners are occupied in conducting important learning activities, and (3) assessing the time-series change in the learning topics that learners are interested in and examine. Such formative assessment can dynamically provide learners with dedicated learning support that is well tailored to each individual learner, and will refine the conventional methodology for supporting real-world learning.

C. Research objective

Real-world learners interact with the world, select from a range of possible actions, and then execute an action. The final output of the brain is behavior, and the intellectual state of real-world learners is to some extent physically reflected by their body. Although it is not possible to directly observe

 ISBN: 978-1-61208-229-5

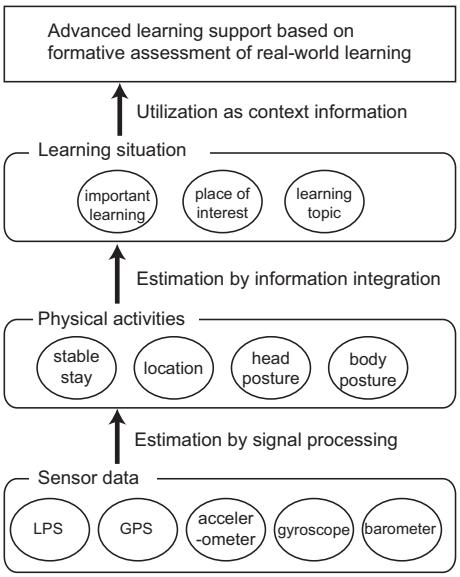


Fig. 1. Estimating learning situation by sensing physical behavior.

the thinking process of learners, their interaction with the world can be observed from outside, and can be a clue to understanding learning situations.

Real-world learners use their body to interact with the surroundings, and in this way obtain knowledge about the world. A learner's body can be considered to be a sensor for obtaining information about the real world, and also an actuator for manipulating the world. Therefore, in the present research, as a fundamental approach for achieving advanced intellectual support based on systematic formative assessment of real-world learning, we developed a sensing technology that can identify the behavior of real-world learners, and estimate the nature of their intellectual activity in real time. The system recognizes physical behavior using low-level signal processing, integrates the recognition results, and thus forms an understanding of learning situations at a high level of abstraction (Figure 1).

D. Table of contents

Section II shows our sensing technique to precisely determine the time series of important learning activities in the world. Section III focuses on the spatial structure of a real-world learning field, and demonstrates that knowledge is three-dimensionally embedded in the world. Section IV proposes our sensing technology to determine the learning topic that a learner is examining, and to understand how learners access knowledge that exists in the real world. Section V discusses our contribution and potential applications. Section VI concludes this paper with a brief summary of our achievements.

II. TIME SERIES OF IMPORTANT LEARNING ACTIVITIES

A. Important learning activities and unimportant activities

Determining how real-world learners carry out important learning activities is useful for forming an understanding of how they access knowledge embedded in the world. Based on past research [4], we can divide environmental learning activities into two categories: important learning activities and unimportant activities. We can define important and unimportant activities as follows:

- Important learning activities
  - **Observation**: An activity for acquiring knowledge through interaction with the world (e.g., observation or survey focusing on a certain target, such as touching plants and soil, or writing field notes while inspecting an observation target).
  - **Knowledge exchange**: A conversation activity for interactively solving a problem through externalizing and exchanging knowledge (e.g., cooperative thinking through conversation and discussion). Trivial chat is not regarded as a knowledge exchange.
  - **Intellectual investigation**: The most important activity, involving simultaneous observation and knowledge exchange (e.g., cooperative thinking and discussion through a collaborative field survey). Physical and internal experiences mutually influence each other. Real-world knowledge embedded in a situation [1], [2] is maximally utilized by intellectual investigation.
- Unimportant activities
  Lack of observation, lack of conversation, idle talk, rest, and so on.

B. Determining intellectual activity through real-world behavior

The time when a learner carries out a learning activity is meaningful information for understanding knowledge acquisition. Although human behavior can be estimated from sensing data [5], [6], [7], [8], [9], it is conventionally difficult to estimate the intellectual state of a learner by using sensors. To begin with, sensors are tools for measuring physical quantities (e.g., velocity, acceleration, angular velocity, signal strength) associated with body movement. Conventional research uses such sensors to recognize human physical behavior. A typical example is identifying the location of humans at a given time [5]. Other examples include recognition of daily activities (e.g., walking, ascending stairs) [6], and abnormal activities (e.g., slipping on a wet floor) [7]. This has recently been extended to the ability to sense human activities using consumer devices such as mobile phones [8], [9]. However, few studies have attempted to determine the relationship between the physical behavior and intellectual state of humans. Hence, even if we measure real-world activities using sensors, it is quite difficult to infer the internal intellectual state of the learner. Thus, we conducted field surveys to investigate the relationship between learner behavior and important learning activities, and obtained the following significant results [10].

1) When learners walk carelessly about in an environment, they rarely have enthusiastic discussions and obtain only superficial information about the environment.

 ISBN: 978-1-61208-229-5

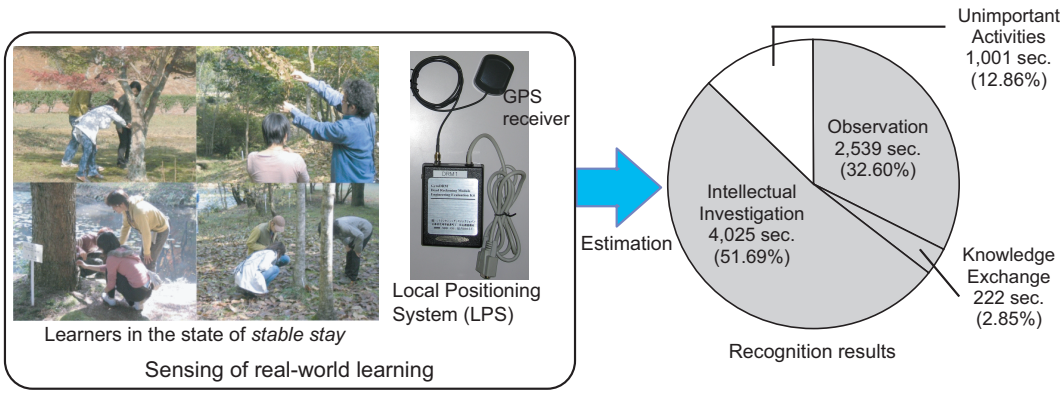


Fig. 2. Automatic accurate identification of important learning activities (precision=87.14%).

2) When learners engage in important learning activities (i.e., observation, knowledge exchange, intellectual investigation), they display a characteristic stay behavior. We call this behavior *stable stay*. Stable stay is defined as a condition that extends for T_t [sec] or more, in which a fixed body posture is adopted (horizontal angular rotation of T_θ [deg/sec] or less) and movement is restricted to a velocity of T_v [m/sec] or less. Stable stay includes the state of crouching down. The learners in Figure 2 are exhibiting stable stay behavior.
3) When learners are not focused on a specific learning topic and casually look around, their body orientation in the horizontal plane is not fixed, even if they are not walking.

C. Recognizing important learning activities by sensing stable stay conditions

Under stable stay conditions, the possibility that important learning occurred was found to be 3.12 times higher than for other conditions [10]. On the grounds that important learning activities and stable stay conditions frequently co-occur [10], each learner was given a wearable local positioning system (LPS) (255 g, 111 x 82 x 39 mm, Figure 2) to be placed on their lower back in order to determine the time series of important learning activities. The LPS is a sensor for recording the local movement and body orientation of a learner. Important learning activities were identified based on the detection of stable stay conditions with three threshold parameters (T_t = 15.00 [sec], T_v = 0.10 [m/sec], T_θ = 60.00 [deg/sec]). These values were determined by pre-evaluating sensor data for two experimental learners and inspecting the data distribution of physical movements associated with both important learning activities and unimportant activities. The data for these two learners were not used in the subsequent experiments for evaluating the recognition accuracy.

We evaluated the accuracy of the recognition method using ground-truth data for the time series of stable stay conditions and important learning activities (18,000 sec; data for five groups representing 15 learners). As shown in the pie chart in Figure 2, the recognition results mostly fell into the important learning category. We confirmed that the method could accurately recognize the time when important learning occurred (precision = 87.14%) by automatically analyzing the LPS data for the learners.

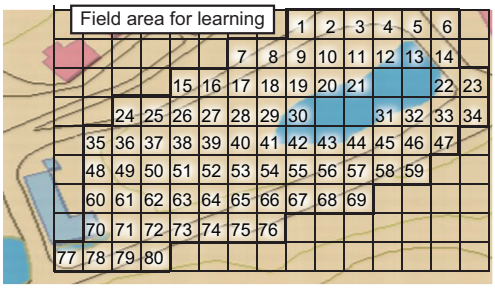


Fig. 3. Region map of Kamigamo Experimental Forest, Kyoto University.

III. KNOWLEDGE THREE-DIMENSIONALLY EMBEDDED IN THE WORLD

A. Regional dependence of learning topics

As shown in Figure 2, learners exhibit a wide variety of behavior such as looking at various parts of the environment, touching objects at various points with their hands, and observing objects while crouching down. However, whatever each learner actually notices, observes or investigates, stable stay behavior is commonly displayed when important learning activities are performed. This is an important finding concerning the common structure of diverse learning activities.

Important learning is a learning activity for accessing knowledge embedded in the real world. By sensing stable stay conditions, an accurate estimation can be made of how such important learning occurs in a learning field. Moreover, by tracing the location of learners with a GPS receiver, both the places of interest and the time spent engaged in important learning activities can be determined. Although this is an advanced approach to determining the time sequence of important learning activities, it is still difficult to estimate what a learner actually learns.

 ISBN: 978-1-61208-229-5

Fig. 4. Learner behavior in a natural ecosystem with a multi-layered hierarchy.

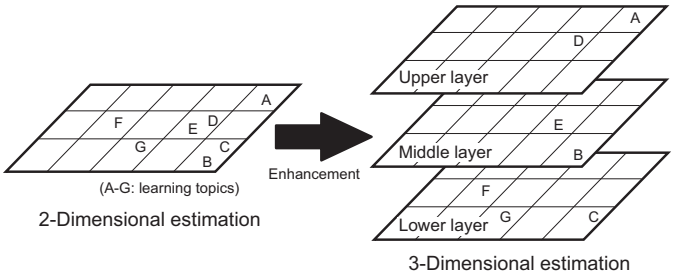


Fig. 5. 3D determination of the learning topics that a learner examines.

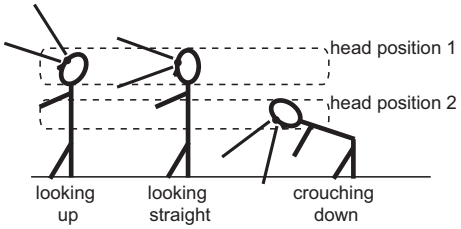


Fig. 6. Observation target and posture.

By identifying the targets on which important learning activities are focused, we hope to be able to clarify the associated content and the knowledge acquisition process. Here, if "what is learned" is related to "the position of the learner", the range of learning topics can be narrowed down based on location. To determine whether this is the case, we conducted experiments in a part (130 x 50 m; Figure 3) of Kamigamo Experimental Forest, Kyoto University, with 15 learners in March, 2010. We found that the learners considered a total of 142 topics. These included, for example, "the symbiotic relationship between mushrooms and moss", and "the relationship between pinecone features and the growth environment of pine trees." We divided the field using a 10 x 10 m grid, and defined 80 different regions (Figure 3). We found that the topics that the learners considered tended to depend on their location. Specifically, on average, each topic was considered in 1.71 regions, and 2.16 topics were considered per region. The topics and the locations are closely related, which illustrates the uniqueness of the physical information in each region.

B. Height dependence of learning topics

Based on the regional dependence of the learning topics, two-dimensional (2D) positional information is useful for narrowing down possible learning topics. However, in regions where more than one learning topic exists, 2D place information is insufficient to uniquely identify the topic of interest. We therefore need to consider additional information for the determination of learner context. We note that a natural ecosystem generally has a multi-layered hierarchy. As shown in Figure 4, learners behave differently even if they stay in the same place. They sometimes look up and sometimes crouch down, and their posture changes according to the learning topics that they consider. We therefore performed a 3D classification of the 142 topics learned in the experiment. We found that the topics were not only region dependent but also height dependent and could be categorized into the following three layers:

- **Upper layer**: Objects above the level of the learners' heads (e.g., tall trees, branches and leaves, light sources). The topics in this category are often examined when a learner is looking up.
- **Middle layer**: Objects at eye level (e.g., tree trunks, bushes). The topics in this category are often examined when a learner is looking straight ahead.
- **Lower layer**: Objects at ground level (e.g., mushrooms, moss, soil, undergrowth, aerial roots, mountain stream, pond). The topics in this category are often examined when a learner is looking down or crouching down.

What a learner examines depends on the region, and each region offers only a small number of topics. Moreover, this research also found that each layer of an ecosystem contains different knowledge, and height is therefore another factor influencing the knowledge that a learner can obtain. Even if more than one topic occurs in the same region, we found that height information is useful for distinguishing such topics. If the learning topics cannot be distinguished in a 2D space, it is possible to distinguish them in a higher 3D space (Figure 5).

IV. Determining the Currently Examined Topic

Stable stay is a condition that is defined in a horizontal 2D plane and can be used for robust identification of an important learning activity. As discussed in Subsection III-B, the posture of learners reflects their interests at that time. We therefore extended the technique for recognizing stable stay behavior to also identify the posture involved, thus allowing the 3D position of the topic of interest to be determined.

A. Wearable sensors

Using an eye-mark recorder, it might be feasible to track the gaze direction of a learner. However, such devices are currently large and heavy, and restrict the movement of learners. Furthermore, when learners move, their location and head direction constantly change without constraint. This makes it difficult to use data from an eye-mark recorder for automatic

 ISBN: 978-1-61208-229-5

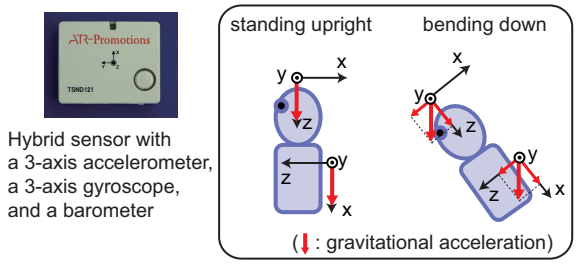


Fig. 7. Gravitational acceleration and learner's posture.

determination of what is being observed. On the other hand, it is known that head direction and gaze direction are closely related, so that head direction can be substituted for gaze direction [11]. For example, as shown in Figure 4 and Figure 6, to observe objects at ground level, learners need to crouch or bend their head down. A learner who is crouching down has a different head height to one who is standing upright (Figure 6). Thus, we classified a learner's posture (e.g., looking straight ahead, looking up, looking down, crouching down) based on head height and body tilt. Each learner was provided with wearable-type hybrid sensors shown in Figure 7 (head and body), in addition to a LPS (lower back) and a GPS (head) for sensing stable stay conditions. The hybrid sensor (22g, 37 x 46 x 12 mm) has a 3-axis accelerometer (50Hz), a 3-axis gyroscope (50Hz), and a barometer (25Hz) built in. The accelerometers are for obtaining 3D tilt information, and the barometers are for obtaining height information. The gyroscopes are for obtaining 3D information on the rotation of a learner's body and head.

B. Estimating posture

When designing a classifier for recognizing human behavior, it is first necessary to understand the typical characteristics of sensor data associated with each of the target behaviors. Thus, we investigated the sensor data obtained in our experiments. As shown in Figure 7, when a learner stands upright and is looking straight ahead, the output from the vertical axes of the accelerometers (i.e., the x-axis of the body sensor, the z-axis of the head sensor) should almost equal the gravitational acceleration (1000 mG). On the other hand, when the body and head of a learner tilts, the accelerometers also tilt. When the body and head rotate about the y-axis, the gravitational acceleration is split into two components along the x-axis and z-axis. For example, as shown in Figure 8, the sensor output exhibits different characteristics when a learner displays different target behaviors (e.g., looking up, standing upright, crouching down). Thus, acceleration data are useful for determining body posture. Data on air pressure and angular velocity are also useful for measuring head position and body rotation.

We are currently in the process of implementing such a body posture identification technique. The integration of multimodal sensor data is carried out using a machine learning method such as a support vector machine (SVM), which can achieve a high generalization performance. The non-linear discrimination function is defined by the below formula.

$$f(\phi(\boldsymbol{x})) = \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x_i}) + b \tag{1}$$

$K(\boldsymbol{x}, \boldsymbol{x_i})$ is a kernel function, and we used the following Gaussian kernel:

$$K(\boldsymbol{x}, \boldsymbol{x_i}) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{x_i}||^2}{\sigma^2}\right) \tag{2}$$

C is the soft margin parameter, and α are Lagrange multipliers. Machine learning in a feature space is carried out by solving the following optimization problem under constraints:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x_i}, \boldsymbol{x_j}) \tag{3}$$

$$Subject\ to \quad 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{4}$$

V. Discussion

Our essential interest centers on understanding how knowledge is generated and develops in the real world. The aim of the present research is to make it possible to observe the process by which intellectual activities occur and progress in a real-world learning field, and to assess such activities using an objective standard. Although recognition of physical behavior can be approached using a machine learning method, it is conventionally difficult to estimate the corresponding intellectual state. This is because there is a wall between understanding superficial-level behavior and the internal-level learning situation. An essential difficulty is that the correlation between physical behavior and the intellectual state is not known.

In order to overcome this difficulty, we have proposed a new approach for sensing intellectual behavior by the complementary use of information on human activity and the spatial structure of the learning field. First, we noted the role of a learner's body as an interactive medium for obtaining knowledge from the surroundings. Thus, we investigated a characteristic stay behavior that can effectively identify situations during which a learner is engaged in important learning activity. We also showed a method for automatically recognizing the occurrence of such behavior using signal processing.

Second, this research confirmed the 2D regional dependence of learning topics in a real-world learning field, and showed that each region offers only a small number of topics. The research also focused on the 3D spatial structure of the field, i.e., the vertical distribution of an ecosystem in a multi-layered hierarchy. The results indicated that each layer contains different knowledge, and height is an important factor influencing the knowledge that a learner can obtain. Even if more than one topic occurs in the same region, we found that height information can be used to uniquely distinguish the topic being learned. We also consider the possibility of classifying human

 ISBN: 978-1-61208-229-5

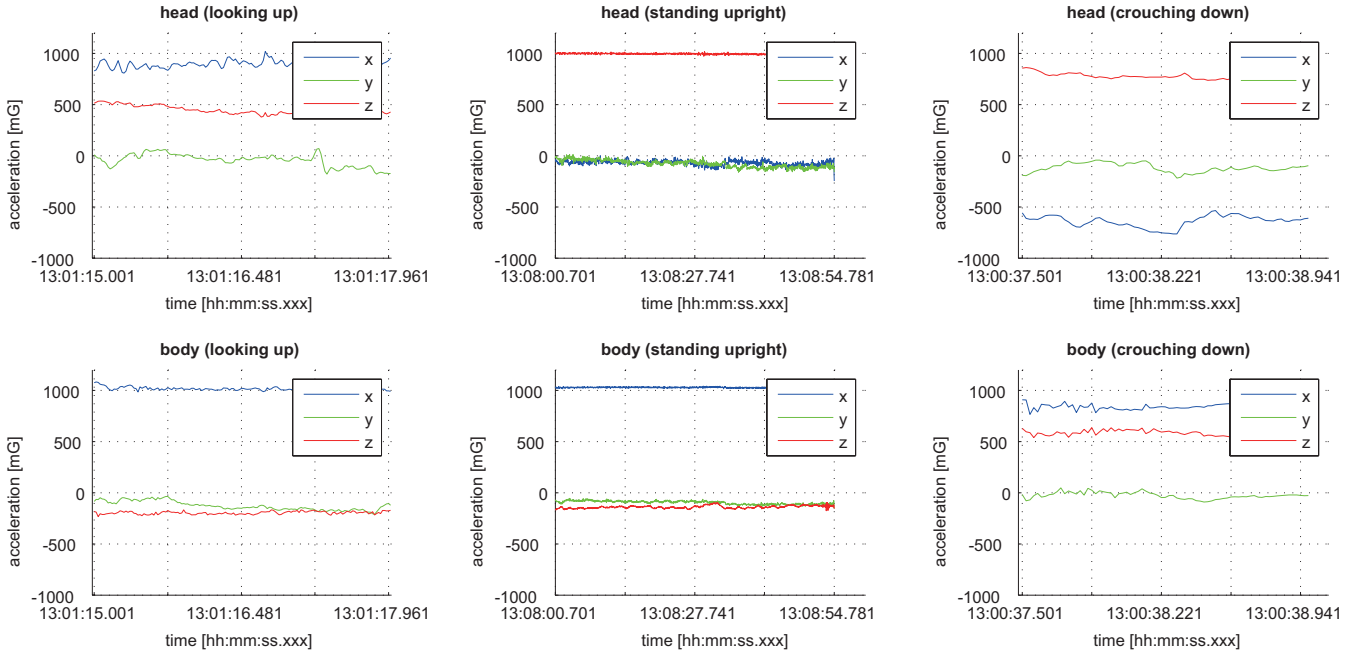


Fig. 8. Learner acceleration data for different target behaviors.

posture with respect to the vertical direction using a machine learning technique to three-dimensionally identify the learner's topic of interest.

This paper examined relationship among the human body, intelligence, and the environment by investigating real-world learning as a model case. Although many studies have been carried out on groupware, learning support, and ubiquitous computing, it is an unsolved and important problem to estimate intellectual situations at a high level of abstraction by integrating the results of low-level signal recognition.

It is hoped that our approach will act as a core for realizing an advanced service for context-aware learning support. For example, it could identify learning objects whose existence is overlooked or whose value is not discovered. This technique could identify 3D locations where knowledge has not yet been found, and could selectively encourage a learner to be aware of potential information. Moreover, numerical indices could be generated of how diversely and actively a learner studied, which are useful for assessing intellectual activity in real-world learning fields.

VI. Conclusion and Future Work

To enhance human intellectual activity, it is important to understand how humans acquire and process knowledge. This paper focused on people who interact with their surroundings and learn from them, and a method was proposed for understanding how they access knowledge that exists in the real world. It involves automatically sensing particular stay behavior that occurs during important learning activities. In addition to the regional dependence of learning topics, it also takes into account the spatial structure of the learning field, i.e., the vertical distribution of an ecosystem in a multi-layered hierarchy. We found that each layer of an ecosystem contains different knowledge, and height is an important factor influencing the knowledge that a learner can obtain. We found that height information can be used to uniquely distinguish the topic being learned. We also discussed the possibility of identifying the learning topic of interest based on the body posture of the learner. Our challenge is to develop a method of estimating the intellectual state of the learner at a high level of abstraction using low-level signal data, and integrating the recognition results.

The rapid progress being made in engineering technologies has led to new ways of innovating learning support. Our intention is not to select engineering technologies that fit conventional methodologies, nor to replace old educational tools with new ones. Our challenge is instead to create an effective method of learning support that can be embodied only using the innovation of computational power. Our research is fundamental for achieving a practical understanding of the dynamics of human knowledge processing and promoting new intellectual activity in a next-generation ubiquitous society.

Acknowledgements

The authors thank the staff at Kamigamo Experimental Forest, Kyoto University, who supported our experiments. This research was funded by a Grant-in-Aid for Young Scientists (B) (22700121) of MEXT.

 ISBN: 978-1-61208-229-5

REFERENCES

[1] J. S. Brown, A. Collins, and P. Duguid, "Situated cognition and the culture of learning," *Educational Researcher*, vol. 18, no. 1, pp. 32–42, 1989.

[2] J. Lave and E. Wenger, *Situated Learning: Legitimate Peripheral Participation*. Cambridge, UK: Cambridge University Press, 1991.

[3] T. Okamoto, N. Nagata, and F. Anma, "The knowledge circulated-organisational management for accomplishing e-learning," *Knowledge Management & E-Learning*, vol. 1, no. 1, pp. 6–17, 2009.

[4] T. Mizukoshi and T. Kihara, *Creation of New Environmental Education (in Japanese)*. Kyoto: Minerva Syobou, 1995.

[5] O. Türkyilmaz, F. Alagöz, G. Gür, and T. Tugcu, "Environment-aware location estimation in cellular networks," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 139:1–139:9, January 2008.

[6] J. Lester, T. Choudhury, and G. Borriello, "A practical approach to recognizing physical activities," *PERVASIVE2006*, vol. LNCS3968, pp. 1–16, 2006.

[7] J. Yin, Q. Yang, and J. J. Pan, "Sensor-based abnormal human-activity detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 8, pp. 1082–1090, August 2008.

[8] N. Györbíró, A. Fábián, and G. Hományi, "An activity recognition system for mobile phones," *Mobile Networks and Applications*, vol. 14, no. 1, pp. 82–91, February 2009.

[9] A. Kobayashi, S. Muramatsu, D. Kamisaka, T. Watanabe, A. Minamikawa, T. Iwamoto, and H. Yokoyama, "Shaka: User movement estimation considering reliability, power saving, and latency using mobile phone," *IEICE Transactions on Information and Systems*, vol. E94-D, no. 6, pp. 1153–1163, 2011.

[10] M. Okada and M. Tada, "Multimodal analysis of spatial characteristics of a real-world learning field," in *Proceedings of 2012 Seventh IEEE International Conference on Wireless, Mobile and Ubiquitous Technology in Education (WMUTE2012)*. Kagawa, Japan: IEEE, March 2012, pp. 25–32.

[11] M. Tada, H. Noma, A. Utsumi, M. Okada, and K. Renge, "Automatic evaluation system of driving skill using wearable sensors for personalized safe driving lecture," in *Proceedings of the IADIS International Conference Mobile Learning 2012*, I. A. Sánchez and P. Isaías, Eds., Berlin, Germany, March 2012, pp. 173–180.

 ISBN: 978-1-61208-229-5

Multiscreen-based Gaming Services using Multi-view Rendering with Different Resolutions

Sung-Soo Kim and Chunglae Cho
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, South Korea
{*sungsoo, clcho*}*@etri.re.kr*

***Abstract*—We present a novel multiscreen-based gaming service system which supports multiple-viewpoint rendering with different resolutions for visualizing a 3D game scene dataset at the same time. Our approach is based on multi-view rendering and reduces the computation to generating video streams for cloud-based gaming services. In addition, the performance speedup of our rendering system is achieved by utilizing both multicore CPUs and a GPU simultaneously without additional requirement for any special hardware. The experimental results demonstrate our multi-view rendering method can be successfully applied to the multiscreen services for the multiplayer games.**

Keywords-multiscreen services; gaming on demand; multi-view rendering; video encoding; video streaming.

I. INTRODUCTION

Consumers desire to access rich multimedia and realistic 3D game content via smartphone, PCs, netbooks, tablets anytime and anywhere. *Multiscreen services* have emerged as a consequence of this user requirement. Users can watch multimedia and game content from any source on any screen through the the multiscreen services. Also, the growth in connectivity and capacity of broadband networks have enabled new forms of *cloud computing*, where data and processing on a remote server is acted upon on a local computing device. This computing model allows a performance focus at a single location, the cloud server, and enables user mobility and pervasive access for the users.

One of the latest advancements in gaming technology that enables such a ubiquitous gaming is *cloud-based gaming services*, also called *Gaming on Demand* (GoD) [1]. Cloud gaming will liberate games from their limiting dependence on consoles, without sacrificing realism, speed, or any other aspect of the true gaming experience. This is a platform-as-a-service approach, analogous to video on demand, where players interact via streamed content generated on the game operator's server rather than players' local systems. There are a number of commercial GoD systems that have been presented to the market such as OnLive [2] and Gaikai [3]. We have idenified four key requirements of cloud-based service systems to provide convincing multiscreen-based gaming services [4]. They are *user responsiveness*, *high-quality video*, *quality of services* and *operating costs*.

Our contributions: We present a novel system architecture for the multiscreen gaming services, which utilizes parallel commodity processors, multi-core CPUs. We also present a novel multi-view rendering algorithm to efficiently support multi-user game on the server, which has a single GPU with multi-core CPUs. In addition, our approach gives the benefits in terms of *arbitrary focal positions* for viewpoints and better rendering quality over prior parallel multi-view rendering methods [5]. This is one of the important features for the multiplayer games in cloud-based gaming services.

The rest of the paper is organized as follows. Section II shows a brief overview of related work. Section III describes the proposed system architecture for cloud-based gaming services. Section IV shows the performance of the proposed method. Finally, Section V ends the paper with some concluding remarks and perspectives for future work.

II. RELATED WORK

In this section, we give a brief overview of related work on cloud-based gaming platforms and parallel rendering algorithms.

Cloud-based gaming platforms: There is a number of commercial cloud-based gaming platforms that have been presented to the market. The *Games@Large* framework enables commercial video game streaming from a local server to remote end devices in local area networks (LANs) [6]. This system and streaming protocols are developed and adapted for highly interactive video games. OnLive is a gaming-on-demand entertainment platform, which their service is available in USA, UK and Belgium [2]. The hardware used is a custom set up consisting of OnLive's proprietary video compression chip as well as standard CPU and GPU chips. Gaikai is developing and delivering a cloud technology platform to put games where they have never been before, including digital TVs, tablets, smartphones, Facebook, and embedded directly into websites. Recently, NVIDIA introduced the *GeForce GRID platform* for gaming-as-a-service providers [7]. The key technologies of this platform are NVIDIA GeForce GRID GPUs with dedicated ultra-low-latency streaming technology and cloud graphics software.

Parallel rendering: Recent work in this area has been focused on *video encoding* and *streaming* techniques to reduce the latency in games [8]. Most of the earlier systems were serial in nature and designed for a single core or processor in terms of 3D rendering. However, the recent trend in computer architecture has been toward developing parallel commodity processors, including multi-core CPUs and many-core GPUs.

 ISBN: 978-1-61208-229-5

It is expected that the number of cores would increase at the rate corresponding to Moore's Law. Based on these trends, many parallel game engines and parallel rendering algorithms have been proposed for commodity parallel processors [9]. Dual-core and quad-core CPU chips are currently available, with some motherboards supporting multiple such chips. Parallel computing is quickly becoming mainstream in the development of computation-intensive applications related to the realistic 3D rendering [10].

OTOY [11] provides technologies that move processor intensive experiences into the cloud; computer applications, operating system, video games, high-definition media content, film/video special effects graphics - fully interactive, in real time, through the power of *server side rendering*. They can deliver the high-quality media content via an interactive stream to any internet enabled device for multiscreen-based gaming services, including PC, iPhone, iPad or TV set top box.

A parallel multi-view rendering architecture in a cluster of GPUs has been proposed in [5]. This system have shown a theoretical analysis of speedup and scalability of the proposed multi-view rendering. However, the critical limitation of this method is that all the cameras are always looking to the center of arbitrary tile. Therefore, this method is not suitable for common mutli-user game applications. Moreover, it is difficult to apply this method to a *high visual quality* games since they used a simple phong shader for lighting and shading.

III. System Architecture

In this section, we describe the proposed system architecture for cloud-based gaming services. Our system consists of three major systems such as distributed service platform (DSP), distributed rendering system (DRS) and encoding, QoS and streaming system (EQS), as shown in Figure 1. The DSP is responsible for launching the game processes on the game execution nodes or rendering job on the DRS after client-side invocation, monitoring its performance, allocating computing resources and managing user information. And, the DSP handles user's game input via UDP from the client-side devices. In client-side, the user's game input is captured and transmitted via UDP by the user input capturing and transmission software on the client devices. Also, the DSP performs execution management of multiple games. In order to perform streaming the game A/V streams to the clients, the DSP requests capturing rendered frame buffer for video encoding and streaming to the EQS.

To improve 3D rendering performance of the DRS, we utilize the multi-threaded game engine [9] that is designed to scale to as many processors as are available within a platform. In order to provide the cloud-based gaming service, the DRS has common system interfaces to the DSP and the EQS. The EQS is responsible for audio/video encoding and streaming the interactive game content to the clients. We use the DirectShow SDK to implement the visual capturing of the games rendered from the DRS. We utilize the H.264 video coding standard for low-delay video encoding of the captured game content. Before the EQS performs the H.264 encoding, we perform a color space conversion from RGB to YUV on the captured frames. Finally, we exploit the Real Time Protocol (RTP) packetization to transmit the encoded video stream in real-time [12].

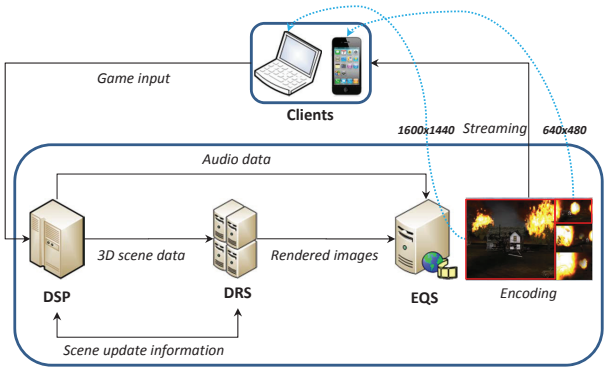


Fig. 1. Our system processing flow: *DSP-Distributed Service Platform, DRS-Distributed Rendering System, EQS-Encoding, QoS and Streaming System*

A. Multi-view Rendering with Different Resolutions

The DRS consists of four major block components such as *rendering scheduler*, *multi-view manager*, *rendering task manager* and *renderer library*. The rendering scheduler is responsible for rendering process monitoring, performance timer control, rendering statistics management and communicating other modules for external rendering requests in the DRS blocks. The key performance improvement for the game applications is the use of per-thread task queues. This eliminates the synchronization checkpoint when one shared task queue is used. Advanced task schedulers may use heuristics to determine which thread to steal from and which task to steal and this may help cache performance. In order to implement the rendering scheduler, we use the Intel Threading Building Blocks (TBB) [9], which is highly optimized scheduler. The multi-view manager is responsible for performing the management of user's viewpoints (such as insertion, deletion, update and search operations) for the shared spaces in the multi-user games. The rendering task manager module performs the rendering task decomposition and parallelization in order to improve the rendering performance. In our work, we use the Object-oriented Graphics Rendering Engine (OGRE) [9], which performs a 3D scene graph management and rendering. In the case of multiscreen-based gaming services, multiple viewpoints for different resolutions are needed if we want to support several users visualizing a given 3D scene at the same time. However, rendering multiple views with different resolutions using the standard graphics pipeline is a challenging problem. In order to provide the interactive multi-view rendering results for the multiscreen-based gaming service, we utilized the shared resources for the rendering such as scene graph, textures and shaders in a GPU as much as possible and keeping the quality of the rendering results. We exploit a video streaming approach to provide the game's encoded video with different resolutions at the same time

 ISBN: 978-1-61208-229-5

Fig. 2. The result of multi-view rendering with different resolutions (1: 1600x1440, 2,3,4: 640x480)

according to the requests of client devices such as PCs, laptops (e.g., 1600x1440 resolution) and smartphones (e.g., 640x480 resolution) as shown in Figure 2.

If R_i denotes a i-th rendered image in framebuffer of the DRS, then $\mathcal{S}_k$, which has i image sequences is defined as:

$$\mathcal{S}_k = \{R_1, R_2, ..., R_i\}$$

The $\mathcal{CP}_i$ denotes the i-th viewpoint parameters, which contains internal parameters such as focal length $f_l(f_x, f_y)$, center $c(c_x, c_y)$, aspect ratio a and external parameters such as position $p(c_x, c_y, c_z)$ and orientation $r(r_x, r_y, r_z)$. The DSP generates this $\mathcal{CP}_i$ and the resolution of the client screen, (x_r, y_r), according to the requests of the clients.

Algorithm 1 Viewpoint addition algorithm.

```
1: procedure ADDVIEW(U_i, CP_i, x_r, y_r)
2:     RenderWindow W;
3:     Camera C_i;
4:     Viewport V_i;
5:     RenderedFrameBuffer R_i;
6:     C_i ← createCamera(U_i, CP_i);
7:     V_i ← addViewport(C_i, x_r, y_r);
8:     R_i ← renderOneFrame(W, V_i, C_i);
9:     return R_i
10: end procedure
```

The DRS provides the function for adding the multiple viewpoints to support the multi-view rendering. First, the DSP receive the service requests from the clients. These requests include several user information, U_i, such as user identification, selected game, which they want to play and initial or previous viewpoints in the 3D game space. Then, the DSP sends these information to the DRS to request for multi-view rendering. According to this request, the DRS provides the function for adding viewpoints, $\mathcal{CP}_i$. To perform this function on the DRS, we create the cameara C_i and viewport V_i objects to attach the viewport to the render window W_i. After the viewport was successfully added to the render window, the DRS performs the rendering procedure to generate an image on the framebuffer in a GPU. The pipeline of our algorithm for multi-view rendering with different resolutions is shown in **Algorithm 1**.

If EA_i and EV_i denote a i-th encoded audio and video in interactive game content respectively, then $\mathcal{ES}_k$, which has i encoded audio/visual gaming sequences is defined as:

$$\mathcal{ES}_k = \{(EA_1, EV_1), (EA_2, EV_2), ..., (EA_i, EV_i)\}$$

Therefore, the EQS performs the streaming $\mathcal{ES}_k$ to the clients for the cloud-based gaming services. In order to address the game's audio/visual output capturing, we develop the capturing module on the EQS in C++ and DirectShow SDK. We also develop the H.264 encoder for achieving low-delay video coding. On the other hand, the client side devices for our system support the H.264 decoding functionality. Also, the client is responsible for capturing the commands of the input controller such as keyboard and mouse, and sending them to the DSP via UDP.

IV. Experimental Results

This section presents the performance results of multi-view rendering with different resolutions and video encoding performed using our method. We have evaluated the performance of multiview rendering on a PC running Windows 7 operating system with Intel Core i7 2.93GHz CPU, 8GB memory and a ATI Radeon HD 5770. We used OGRE library based on DirectX as a graphics API and Microsoft HLSL for a shading language. The frames per second (FPS) is the number of frames per second that have been rendered by the DRS. High FPS results with smooth movements in the 3D scene.

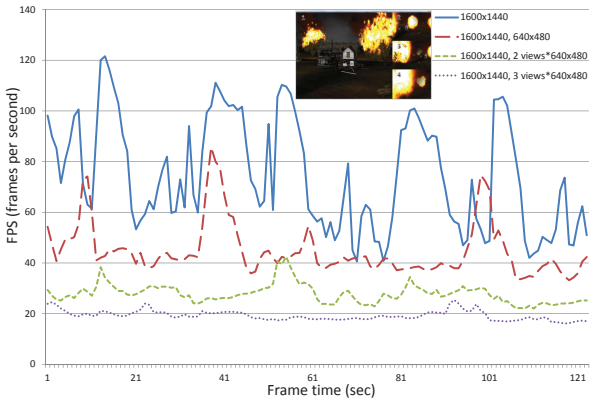


Fig. 3. The performance of multi-view rendering with different resolutions.

Our system rendered the single view (1600x1440) at 74.7 fps on average with one GPU. We measured the FPS at the DRS for multi-view rendering with 1600x1440 and 640x480 resolutions. In the case of multi-view rendering with a 1600x1440 view and two 640x480 views at the same time, we can get 27.8 fps on average with one GPU. Figure 3 shows the performance result of multi-view rendering according to the resolutions. In addition, in terms of scaling performance according to the number of CPU cores, our rendering performance using multi-core CPUs (8-core) achieves 4.2x speedup

 ISBN: 978-1-61208-229-5

over execution using single-core CPU. In order to analyze our video coding performance for the use of streaming game output to client devices, we have performed the experiments using H.264 codec. Our encoding system can encode in 25.6ms on average for eight views (interactive gaming videos) with 640x480 resolutions at 15 fps in parallel. Also, our encoding processing time is 23.1ms on average time per frame (1600x1440) as shown in Figure 4. The output of the first-person shooter game using the Unreal Development Kit (UDK) was captured and encoded.

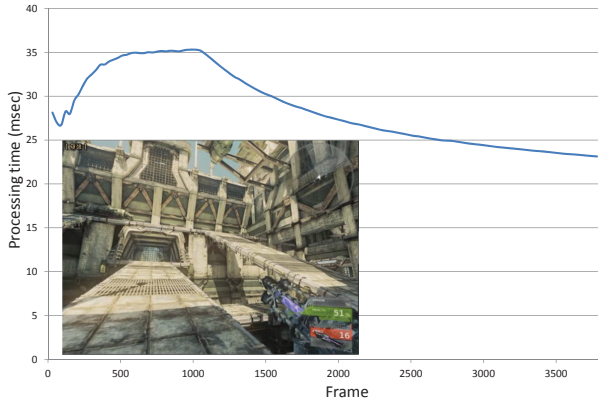


Fig. 4. The performance of video encoding for 1600x1440 at 30 fps.

Analysis: Our rendering system provides good performance scaling of multi-core CPUs for multi-view rendering with different resolutions and video encoding. And the multi-view rendering algorithm maps well to the current GPUs and we have evaluated its performance with different rendering resolutions. Compared to the prior parallel multi-view rendering method [5], our approach offers the advantage for multi-user games by supporting various viewpoints with *arbitrary focal positions*. Our algorithm can easily handle insertion and removal of viewpoints with different resolutions and can also take advantage of scalable and parallel processing using multi-core CPUs. Furthermore, it is relatively simple to combine the video encoding methods and optimizations in the cloud-based gaming platform. This makes it possible to develop a more flexible GPU-based framework for the video encoding methods like H.264/AVC.

Our approach has some limitations. First, we support the multi-view rendering for one multi-user game, since it is difficult to share the rendering resources in a GPU among different games. We believe that this can be resolved by using multi-GPUs. Secondly, our system performs directly rendering to the framebuffers on the server-side machines. However, in terms of efficient services in the cloud-based gaming, we should exploit the *off-screen rendering* approaches and *GPU virtualization* techniques [7].

V. Conclusion and Future Work

In this paper, we have presented the system architecture for the multiscreen-based gaming services and multi-view rendering with different resolutions. The performance speedup of our rendering system is achieved by utilizing both multicore CPUs and a GPU simultaneously without additional requirement for any special hardware. We found that the proposed system provide the multi-view rendering for different focal positions for each viewpoint with high visual quality. In addition, we demonstrate that the proposed rendering system could prove to be scalable in terms of parallel rendering. So, we believe that our rendering system will provide high-quality with good performance for the multiscreen-based gaming services.

There are many avenues for future work. It is possible to use new capabilities and optimizations to improve the performance of the video encoding especially H.264/AVC through the GPU-based implementation. Furthermore, we would like to develop algorithms for integrating the multi-view rendering with the video encoding in a GPU.

Acknowledgments

The game technology demo (Intel's smoke demo) in Figure 2 is courtesy of the Intel Corporation. This work was supported in part by the IT convergence R&D program of the Ministry of Knowledge Economy (MKE)/KEIT [10039202], *Development of SmartTV Device Collaborated Open Middleware and Remote User Interface Technology for N-Screen Service*.

References

[1] T. Karachristos, D. Apostolatos, and D. Metafas, "A real-time streaming games-on-demand system," in *Proceedings of the 3rd international conference on Digital Interactive Media in Entertainment and Arts*, ser. DIMEA '08. New York, NY, USA: ACM, 2008, pp. 51–56. [Online]. Available: http://doi.acm.org/10.1145/1413634.1413648

[2] S. Perlman. (July 2012) Onlive launches in belgium. [Online]. Available: http://blog.onlive.com/2012/07/30/onlive-launches-in-belgium/

[3] G. website. (2012) What is gaikai? [Online]. Available: http://www.gaikai.com/

[4] S.-S. Kim, K.-I. Kim, and J.-H. Won, "Multi-view rendering approach for cloud-based gaming services," in *The Third International Conference on Advances in Future Internet*, ser. AFIN 2011, 2011, pp. 102–107.

[5] W. Lages, C. Cordeiro, and D. Guedes, "A parallel multi-view rendering architecture," in *Proceedings of the 2008 XXI Brazilian Symposium on Computer Graphics and Image Processing*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 270–277. [Online]. Available: http://portal.acm.org/citation.cfm?id=1440461.1440894

[6] Y. T. A. S. F. Bellotti and A. Jurgelionis., "Games@large - a new platform for ubiquitous gaming and multimedia," in *Proceedings of the Broadband Europe Conference*, ser. BBEurope '06, 2006, pp. 11–14.

[7] P. Eisler. (2012, June) What to expect from geforce grid for cloud-based gaming. [Online]. Available: http://blogs.nvidia.com/2012/06/what-you-can-expect-from-geforce-grid/

[8] A. Jurgelionis, P. Fechteler, P. Eisert, F. Bellotti, H. David, J. P. Laulajainen, R. Carmichael, V. Poulopoulos, A. Laikari, P. Perälä, A. De Gloria, and C. Bouras, "Platform for distributed 3d gaming," *Int. J. Comput. Games Technol.*, vol. 2009, pp. 1:1–1:15, January 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/231863

[9] J. Andrews. (June 2009) Designing the framework of a parallel game engine. [Online]. Available: http://software.intel.com/en-us/articles/designing-the-framework-of-a-parallel-game-engine/

[10] S. Eilemann, M. Makhinya, and R. Pajarola, "Equalizer: A scalable parallel rendering framework," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, pp. 436–452, May 2009. [Online]. Available: http://portal.acm.org/citation.cfm?id=1515609.1515684

[11] OTOY. (August 2012) Otoy website. [Online]. Available: http://www.otoy.com/

[12] R. 3550, *RTP: A Transport Protocol for Real-Time Applications*.

 ISBN: 978-1-61208-229-5

Examining User Intention Behaviour towards e-Readers in Japan Using the Decomposed Theory of Planned Behaviour

Qazi Mahdia Ghyas, Hirotaka Sugiura, Fumiyo N. Kondo
Dept. of Social Systems & Management
University of Tsukuba
Ibaraki, Japan
e-mail: s1030160@u.tsukuba.ac.jp, sugiura80@sk.tsukuba.ac.jp, kondo@sk.tsukuba.ac.jp

Abstract— **The e-reader is a mobile electronic device designed specifically for reading electronic books. E-readers have captured public attention all over the world, making it essential to better understand the patterns of user adoption and intention behaviour regarding these devices. In this paper, we examine two adoption models, the Theory of Reasoned Action (TRA) and the Theory of Planned Behaviour (TPB), and four Decomposed Theory of Planned Behaviour (DTPB) models with our proposed extended antecedents (hedonic and utilitarian) and determine that the DTPB model-2 is, relatively speaking, the best among all of these models. In terms of the Akaike information criterion (AIC) and the Browne–Cudeck criterion (BCC), the TRA is more accurate than the TPB. However, the other fit index, the Root Mean Square Error of Approximation RMSEA (RMSEA) is not acceptable for the TRA and the TPB (> 0.1) although it is for the DTPB models. As e-readers offer an increasing variety of products to use (e-books, music, applications, etc), this will change users' beliefs regarding the opportunities that are needed to perform a behaviour. In our conclusion, the utilitarian products offering by e-reader are an important variable that influences consumer intention to use e-book reader, but the variable of hedonic product is not.**

Keywords- E-reader; DTPB model; User intention; Type of products.

I. INTRODUCTION

An e-reader is a mobile electronic device designed primarily for the purpose of reading digital e-books and periodicals [1]. The main advantages of the e-reader are convenience and information access. Users have the ability to access information anytime, anywhere, and will appreciate that this access is fast and easy. E-readers are becoming ever more popular in high-tech cultures, such as that in Japan, the US, and Europe. According to the latest research from Informa Telecoms & Media (2012), e-reader sales are expected to peak at 14 million in 2013 [2]. The consumption of e-books is also growing in Japan. Interestingly, 80% of all e-books read in Japan are consumed on mobile devices [3].

As Japanese youth are technology savvy, they typically feel great enthusiasm towards new technologies. However, e-readers are not yet as popular among young Japanese as other new technologies (e.g., smart phones). It has become extremely important for the e-book industry to explore the consumer attitudes and adoption behaviours regarding this technology. Examining Japanese consumer perceptions of e-reading devices is essential for current and future device development.

Our main objectives are, first, to achieve a clear understanding of consumer attitudes towards e-readers; we have investigated the antecedents related to their adoption and usage. Second, we aim to generate a research model that accurately describes Japanese youths' e-reader usage behaviour and belief structure. There is no prior research comparing the TRA [4], TPB [5], and DTPB [6] models regarding e-book usage in this population. To accomplish these objectives, the TRA, TPB, and DTPB models have been used as a guideline, and we have considered both hedonic [7] and utilitarian products [7] as the two extended antecedents of our DTPB models.

One general research question drove this study: how do students' multidimensional beliefs influence their adoption of or intention to use e-readers. In an effort to answer this question, here we examined four different DTPB models. Among them, DTPB model-2 is better in terms of RMSEA and chi-square. It also explains how potential users' intentions are influenced by significant paths of attitude, subjective norm, and perceived behavioural control, regarding their decomposed antecedents. DTPB model-2 was statistically significant for our proposed antecedent of utilitarian products, which improves our understanding of users' perceived behavioural control regarding e-readers; however, the hedonic product was not found to be significant. Users appeared to be more focused on the utilitarian aspects of e-readers rather than their hedonic aspects, rendering the latter statistically insignificant to our results.

This paper is organized as follows: the next section focuses on the literature review and the theoretical background. Then, we conceptualize the research model and propose our hypotheses. The subsequent section describes the research methodology and empirical findings. Finally, the study discusses the implications of the research in terms of theoretical and practical contributions and provides concluding remarks with limitations and future research directions.

 ISBN: 978-1-61208-229-5

II. LITERATURE REVIEW

A. e-Reader market in Japan

In a recent edition of the eBook Journal, Yashio Uemura of Tokyo Denki University [8] laments that the current e-book boom in Japan is in reality a boom in e-book seminars. This sense of frustration within the industry may seem at odds with its annual revenues, as reported by Impress R&D, of $600 million and growth in excess of 20% per year. These impressive numbers belie the fact that comics make up 75% of this revenue and that, apart from comics and magazines, there seems little significant advance in broadening the e-book consumer base within Japan [9]. Currently, there are two e-reader devices in the Japanese market, the Sony Reader and the Biblio Leaf, which is available through the mobile retailer KDDI [10]. It is expected that the increased volume and richness of e-book content will spur the formation of new e-reader markets. As shown in Fig. 1, the e-reader market includes any device that can be used for reading, including tablet personal computers, such as the iPad [4].

B. e-Reader adoption models

One of the important and significant issues related to IT is the identification of factors that cause people to accept new technologies and information systems and to use them [11]. Several relevant theories are offered, such as Theory of Reasoned Action [4], the Technology Acceptance Model (TAM) [12], the Theory of Innovation Diffusion [13], the TPB [5], and the DTPB [6]. In addition, Venkatesh, Morris et al. [14] developed the Unified Theory of Acceptance and Use of Technology (UTAUT) model.

MA Jiah et al. [15] proposed a model that would test the TAM's effectiveness in determining the influence factors on the acceptance and use of e-readers. Sungjoon Lee [116] examined the factors that lead to the adoption of the mobile e-book in South Korea. Jaemin Jung et al. [17] identified the predictors of e-reader diffusion with regard to consumer awareness, interest, and intention to use. Malathi Lectumanan et al. [18] have investigated consumer intentions of using e-books as educational aids by using the TAM. Brown [19] has developed a research framework demonstrating college professors' and students' acceptance of e-books and e-readers as a viable alternative to traditional paper textbooks, as well as their acceptance of these technologies for use in the classroom. Shih-Chun Chou [120] has compared pre-adoption and post-adoption beliefs to determinants of e-reader adoption and continuation. Bram Pynoon et al. [21] examined secondary school teachers' acceptance of a digital learning environment (DLE) using the UTAUT model. Jung-Yu lai et al. [22] has explored the factors that drive users to use dedicated e-readers for reading e-books.

However, no previous research has attempted to understand the behaviour of e-reader consumers by using the TRA, the TPB, and the DTPB. Here, these three models, that is, the theory of reasoned action, theory of planned behaviour, and decomposed theory of planned behaviour, are considered and compared in order to investigate the attitudes of e-reader users. In the next section, I explain the TRA, TPB, and DTPB models.

- Model 1: Theory of Reasoned Action (TRA)

This theory, developed by Fishbein and Ajzen [4], is one of the most important theories used to explain the human behaviours [23]. According to the theory, behavioural intention (to use a technology) is explained by people's attitudes toward that behaviour and subjective norms.

- Model 2: Theory of Planned Behaviour (TPB)

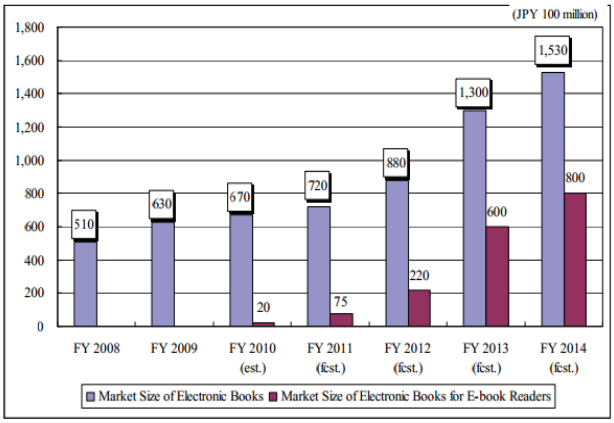
Ajzen [5] developed the theory of reasoned action by adding the construct 'perceived behavioural control' into the model as a determinant of behavioural intention and behaviour. It determines the impact of three factors, which are 'attitude', 'subjective norms', and 'perceived behaviour control' on the tendency to behave in a certain fashion [6].

- Model 3: Decomposed Theory of Planned Behaviour (DTPB)

The Decomposed Theory of Planned Behaviour was developed by Taylor and Todd in 1995 [6], as illustrated in Fig. 2. They developed the theory of planned behaviour through breaking down the structure of attitude, subjective norm, and perceived behavioural control [24]. This provided an increased ability to explain behavioural intentions and enable the accurate understanding of behavioural events [25]. According to the DTPB, individuals' use behaviours *vis-a-vis* information technology are determined by their 'intention to use'. 'Intention to use', in turn, is determined by the attitude toward behaviour, subjective norm, and perceived behavioural control. Perceived usefulness is the extent to which a person believes using a particular technology will improve their job performance [26]. Perceived compatibility is the extent to which an innovation is consistent with the existing values, past experiences, and current needs of potential adopters [23]. Relative advantage occurs when the perceived advantages resulting from the use of a technology exceeds other alternatives [27]. While the theory of planned behaviour simply explains the relationship between the structure of beliefs and the prerequisite of intention, the decomposed theory of planned behaviour offer a comprehensive approach to understanding the factors affecting users' intentions to use information technology [28]. Within this theoretical framework, complexity plays a significant role in the technology adoption decision, while there is a direct relationship between other features of model and behavioural intention [6].

Here, we use the DTPB model instead of the UTAUT model because the UTAUT model attempts to explain the relationship between perceived usefulness, ease of use, and intention to use as modified by age, gender, and experience. This model will be diversified more by demographic factors than belief factors. Additionally, as we wanted to focus on students' behaviours as situated within an immature e-reader market, we did not use the UTAUT model.

 ISBN: 978-1-61208-229-5



(est. =estimate, fcst. =forecast)

Figure 1. Market size of electronic books and E-book readers in Japan.

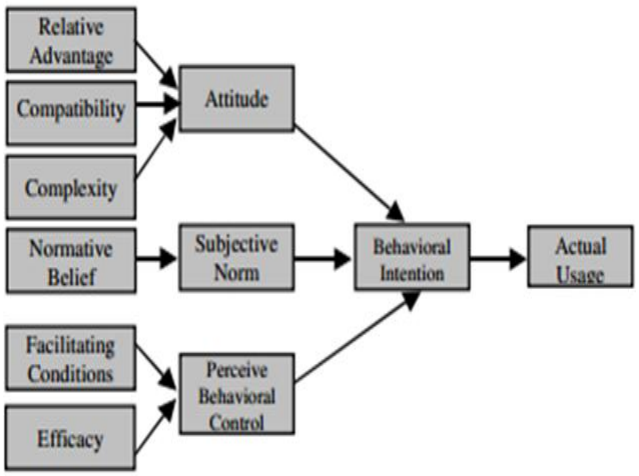


Figure 2. The Basic Decomposed Theory of Planned behavior model.

Our research is an extended version of an unpublished e-reader study [10] that used the basic DTPB model with limited sample size.

C. Our proposed model

With respect to e-reader research, MA Jiah et al [15] have shown in their technology acceptance model that there is a significant relationship among product features and perceived ease of use. Perceived behaviour control is that which refers to the perceived ease or difficulty of performing the behaviour [4]. Sanjukta et al. [7] have shown that the type of product is likely to be an important driver of PBC for internet shopping behaviour. Products can be classified as hedonic or utilitarian. These classifications are primarily intended to better understand how consumers search for, evaluate, choose, take delivery of, and consume different types of products [7]. Both hedonic and utilitarian products offer benefits to the consumer, the former primarily in the form of experiential enjoyment and the latter in practical functionality [7]. The hedonic products offered by e-readers are games, music players, voice recorders, etc. The utilitarian products offered by e-readers are calendar applications, contact list applications, and r eading Essentials (zoom/size, page jump, bookmarks, search, and auto page). Perceived behavioural control (PBC) reflects consumer beliefs regarding access to the resources and opportunities needed to perform a behaviour [6]. The type of product (e.g., e-books, music) featured on e-readers can be an important driver of consumer purchase behaviour [7]. Therefore, we hypothesise that both hedonic and utilitarian products have influence on PBC in the case of e-readers and we add these antecedents to our proposed DTPB models. Figure 3 shows our proposed model with the addition of the new antecedents of hedonic and utilitarian products, along with the basic DTPB model used in previous studies [6], [10].

The same hypotheses (H1–H10) by Taylor et al. [6] and Koeder et al. [10] are included and illustrated in Figure 2. H11 and H12 are our proposed hypotheses for e-readers. The hypotheses are as follows:

H11: That the hedonic products offered by e-readers positively affect perceived behavioural control.

H12: That the utilitarian products offered by e-readers positively affect perceived behavioural control

III. METHODOLOGY

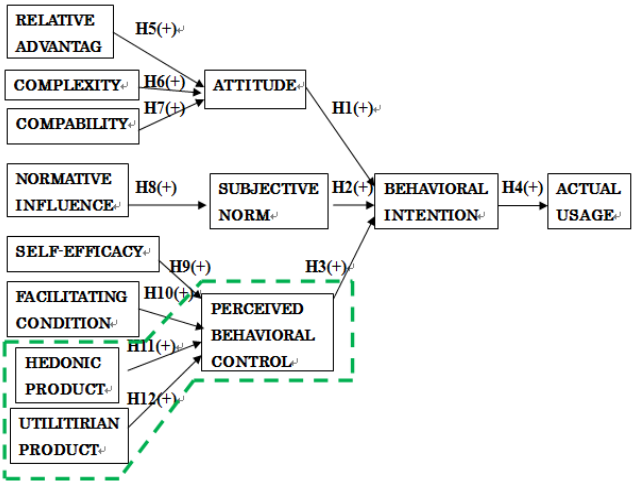
A. Data collection

- Question Development

In developed countries, including Japan, research has been conducted to develop initial models for the adoption and usage of these devices. For our initial qualitative research interviews, the survey items were adapted from previous studies [4], [6], [10] to develop our initial survey instrument. Items to measure behavioural intention, attitude, subjective norm, and perceived behavioural control were based on scales developed by Ajzen and Fishben [4]. Items to measure relative advantage, complexity, and compatibility were based on scales developed by Taylor et al. [6]. Facilitating condition and self-efficacy items were generated based on the work of Ajzen [5]. The survey instrument was pretested with students (N = 13) at the University of Tsukuba, Japan. Based on pretest results, items were revised to ensure reliability and the logical flow of questions. The pretest sample was not included in the final data set.

Behavioural beliefs were adapted from the scale developed by Koeder et al. [10]. The scale included seven items on a 7-point range bipolar-scale: 1 = ‘strongly disagree’ to 7 = ‘strongly agree’. The attitudes toward e-readers were measured using a scale developed by Taylor & Todd [6]. The 7-point semantic differential scales included the following sets: stupid/wise, bad/good, very bad/very good, very unimportant/very important. The other items were measured using a 7-point range bipolar scale (1 = ‘less likely’ to 7 = ‘more likely’) adapted from Taylor and Todd [6]. Consumer demographic characteristics were measured for a descriptive

 ISBN: 978-1-61208-229-5



(+ means positive relation)

Figure 3. Our proposed Decomposed Theory of Planned Behavior model with new hypotheses.

purpose and included gender, age, occupation, income, and reading habits.

- Subjects

Participation in the survey was voluntary and the survey instrument was developed using the software Qualtrics and was administrated online. The English version of the questionnaire was translated to Japanese. This has been checked to ensure its accuracy. The students of the University of Tsukuba, Japan, were selected as the population of interest. Students comprised our sample because it is mainly students who are, or will be, the primary users of e-readers for reading, convenience, and information access. In an effort to determine the intention of students to adopt e-readers as tools, a survey was conducted towards the end of 2011. With a total of 164 completed responses obtained, 151 usable data sets were analyzed to describe the hypothesised paths using structure equation model.

IV. RESULTS

A. Descriptive statistics

A profile of the respondents who participated in this study is displayed in Table 1, which shows that 24.4% of the responding students were female and 75.6% were male.

B. Reliability and validity

We have tested the Cronbach α coefficient for our research, a reliability coefficient of the measured value of questionnaire items for each construct from the point of view of internal consistency, which is used to verify whether each item shows common parts. If the coefficient is 0.7 or more, the internal consistency of the measurement scale is considered to be high and its reliabilities are adequate. The coefficients for each factor are shown in Table 2. Since all values exceed 0.7, the items in this intention model are judged to have shown common parts [29], which render the model acceptable for exploratory research [30]. Test items with lower values than 0.7 have been excluded.

TABLE I. DEMOGRAPHIC CHARACTERISTICS OF THE RESPONDENTS

Gender	Total	%
Men	124	75.6
Women	40	24.4
sum	164	100.0

Dept. of the students	Total	%
Social technology	144	95.4
other	7	4.6
sum	151	100.0

TABLE II. CRONBACH'S ALPHA FOR EACH CONSTRUCTS

Items	No	Cronbach's Alpha
Behavioral Intention	3	0.822
Actual Usage	2	0.711
Attitude	2	0.925
Subjective Norm	3	0.843
Perceived Behavioral Control	3	0.729
Relative Advantage	3	0.833
Compatibility	3	0.847
Complexcity	3	0.850
Normative Influence	3	0.876
Self-Efficacy	3	0.788
Facilitating condition	2	0.709
Utilitarian Product	3	0.809
Hedonic Product	2	0.930

Confirmatory factor analysis (CFA) is a visual representation that specifies the model's constructs, indicator variables, and interrelationships. CFA provides quantitative measures of the reliability and validity of the constructs. In order to check the properties of the measurement scales, we conducted CFA to assess reliability, convergent validity, and discriminant validity. In order to assess the reliability of all the measurement scales, we calculated composite reliabilities (CR) for internal consistency and average variance extracted (AVE) for construct convergence for each construct by using the formula proposed by Fornell et al. [31]. The recommended value of CR is suggested as 0.7 by Hulland et al. [32]. A marginal but acceptable AVE value is 0.4 or higher, as has been reported and used in marketing literature [33], [34], [35]. In addition, we calculated AVE that exceeds the squared intercorrelations (SIC) of the construct with other constructs in the model in order to ensure discriminant validity [32]. CR, AVE, and SIC for each construct of the e-reader are shown in Table 3 where AVE > SIC, AVE > 0.5, and CR > 0.7. Therefore, we may be able to conclude that the reliability of the constructs developed for the e-reader was acceptable.

 ISBN: 978-1-61208-229-5

TABLE III. CR, AVE AND SIC FOR EACH CONSTRUCTS

Convergent Validity	AVE	CR	Discriminant validity	SIC
RA	0.64	0.84	Relative advantage <-> Complexity	0.05
COMP	0.65	0.85	Relative advantage <-> Compability	0.10
COMX	0.68	0.86	Relative advantage <-> Normative influence	0.04
BI	0.60	0.82	Relative advantage <-> facilitating condition	0.06
AU	0.56	0.72	Complexity <-> Compability	0.05
SN	0.67	0.85	Complexity <-> Normative influence	0.43
PBC	0.61	0.80	Complexity <-> facilitating condition	0.33
NI	0.71	0.88	Compability <-> Normative influence	0.09
UTI	0.57	0.80	Compability <-> facilitating condition	0.08
ATT	0.85	0.92	Normative influence <-> facilitating Condition	0.14
FC	0.59	0.73	Normative influence <-> Utilitarian product	0.10
Discriminant validity AVE>SIC Convergent validity AVE>0.5 CR>0.7			Complexity <-> Utilitarian product	0.189
			Compability <-> Utilitarian product	0.25
			Relative advantage <-> Utilitarian product	0.09

C. Model fitting test

Fit statistics, including chi-square, normed fit index (NFI), root mean square error of approximation (RMSEA), goodness of fit (GFI), and adjusted goodness of fit (AGFI), AIC, and BCC were used to assess model fit. An omnibus cut-off point of 0.90 has been recommended for GFI. For GFI, it is generally accepted that values of 0.90 or greater indicate well fitting models [36]. A value of about 0.08 or less for the RMSEA would indicate a reasonable error of approximation [37]. χ^2 to degrees of freedom ratios in the range of 2 to 1 is indicative of an acceptable fit between the hypothetical model and the sample data [38]. With parsimony fit measures, such as the AIC and the BCC, smaller values of these criteria indicate a better fit of the model [39]. In explanation, the total coefficient of determination (TCD) R^2 for the structural equations has shown in this study. The fit statistics and the R^2 values for each are shown in Table 4. These results indicate a preference for DTPB model 2.

- Theory of reasoned action

In Table 4, the statistics indicate that the TRA model provides a measurable fit to the data. χ^2 to degrees of freedom ratio is 3.56, GFI = 0.90, AGFI = 0.82, CFI = 0.90, AIC = 186.87, BCC = 130.55, RMSEA = 0.13. In terms of predictive power, the variance in all dependent variables are R2BI = 0.17, R2AU = 0.32, respectively. The path coefficients are as hypothesised in each case ($p > 0.05$ in all instances). Attitude and subjective norm are a significant determinant of behavioural intention. A further significant determinant of actual use is behavioural intention.

- Theory of planned behaviour

In Table 4, the statistics indicate that the TPB model provides slightly same fit to the data as the TRA fit. However, there is a slight improvement in the fit and the explanatory power of behavioural intention. $\chi^2$to degrees of freedom ratio is 2.61, GFI = 0.88, AGFI = 0.82, CFI = 0.90, AIC = 127.55, BCC = 192.19, RMSEA = 0.1. In terms of predictive power, the variance in all dependent variables is R2BI = 0.20 and R2AU = 0.32, respectively.

The path coefficients are as hypothesised in each case ($p > 0.05$ in all instances). Attitude, subjective norm, and perceived behaviour control are significant determinants of behavioural intention. A further significant determinant of actual use is behavioural intention.

- Decomposed Theory of planned behaviour

We have conducted four DTPB models in order to establish the best-fit index. The four DTPB models are as follows:

DTPB-1 (with FC): DTPB model with the path of facilitating condition →perceived behavioural control.

DTPB-2 (with UTI): DTPB model with the path of utilitarian product → perceived behavioural control. We have tested these two paths individually because they have a correlation of 0.49.

DTPB-3 (with FC and UTI): DTPB model with both paths: facilitating condition → perceived behavioural control; utilitarian product → perceived behavioural control.

DTPB-4 (without PBC): model without PBC construct. Here, this has low R2 = 0.02-0.03 as with some previous studies [10], [40].

Among the four DTPB models, the better models are DTPB-1 and DTPB-2. For model 1, $\chi^2$to degrees of freedom ratio is 2.044, GFI = 0.761, AGFI = 0.713, RMSEA = 0.083, CFI = 0.848, AIC = 796.959, BIC = 960.047, BCC = 798.877, and CAIC = 1023.047. For model 2, $\chi^2$to degrees of freedom ratio is 1.969, GFI = 0.77, AGFI = 0.723, RMSEA = 0.080, CFI = 0.86, AIC = 801.395, BIC = 1009.588, BCC = 834.470, and CAIC = 1078.588. In terms of predictive power, the DTPB-2 explains attitude, subjective norm, and behavioural intention. The variances in all dependent variables are as follows: R2BI = 0.20, R2AU = 0.31, R2ATT = .32, R2SN = .30, R2PBC = .02, respectively for model 2. The low R square value of PBC indicates that utilitarian products and facilitating conditions alone could not provide a powerful explanation of PBC. Mathieson et al. [40] found that PBC did have a significant relationship with behavioural intention, though it did not provide substantial explanatory power.

In Figure 4 and Figure 5, the path coefficients are significantly positive in each case ($p > 0.05$ in all instances) for DTPB-1 and DTPB-2. Attitude, subjective norm, and perceived behaviour control are significant determinants of behavioural intention. A further significant determinant of actual use is behavioural intention. Normative influence is significantly related to SN. Self-efficacy and hedonic products do not significantly and positively influence PBC. However, utilitarian products are significantly and positively related to PBC in model 2 and facilitating conditions are significantly and positively related to PBC in model 1. Taken

 ISBN: 978-1-61208-229-5

TABLE IV. VALUES OF MODEL SELECTION CRITERIA FOR EACH MODEL

Model fit	GFI	AGFI	RMSEA	AIC	BCC	χ^2/df	CFI	R^2_{AU}	R^2_{BI}	R^2_{ATT}	R^2_{SN}	R^2_{PBC}
TRA	0.90	0.82	0.13	127.60	130.60	3.56	0.90	0.32	0.17	N/A	N/A	---
TPB	0.88	0.82	0.10	186.90	192.20	2.61	0.90	0.32	0.20	N/A	N/A	N/A
DTPB-1 (with FC)	0.76	0.71	0.08	796.20	798.90	2.04	0.84	0.31	0.22	0.30	0.30	0.03
DTPB-2 (with UTI)	0.77	0.72	0.08	801.40	834.50	1.96	0.86	0.31	0.20	0.32	0.30	0.02
DTPB-3 with (UTI and FC)	0.73	0.68	0.09	921.30	955.30	2.14	0.81	0.31	0.21	0.30	0.30	0.03
DTPB- 4 (without PBC)	0.82	0.77	0.09	481.60	498.60	2.10	0.89	0.31	0.17	0.30	0.30	---

together, utilitarian products and facilitating conditions do not give a good fit for model 3 as because they have correlation = 0.5. Attitude, subjective norm, and PBC are positively and significantly related to behavioural intention. Behavioural intention is positively and significantly related to actual usage for both DTPB models 1 and 2.

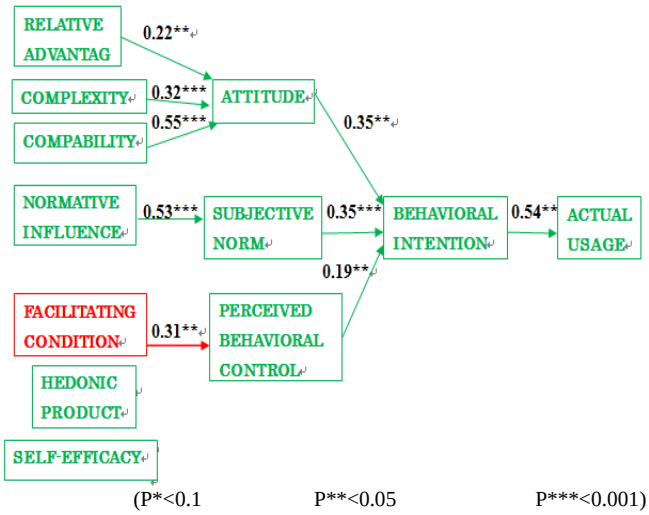


Figure 4. DTPB model 1 with significant paths (same components as Koeder et al.(28))

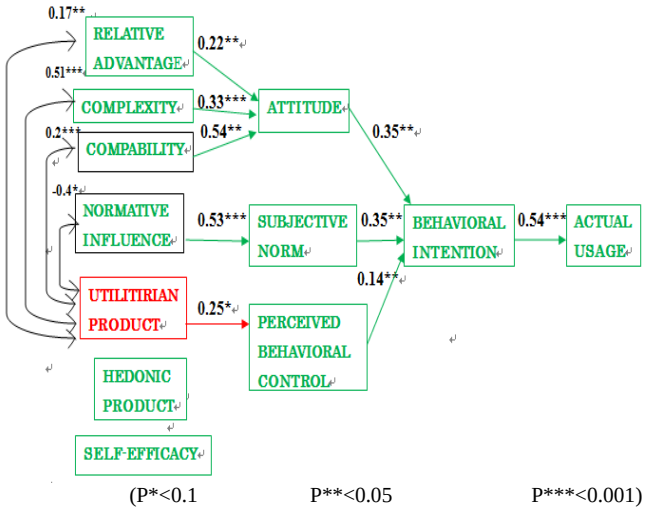


Figure 5. DTPB model 2 with significant paths (our proposed components)

V. DISCUSSION

This study compared the TRA, TPB, and DTPB models with the extension of product characteristics. The aim was to provide useful and interesting results that demonstrate the best model for predicting consumer behaviour with regard to the adoption of e-readers, thus helping e-readers' developers refine their strategic planning and enhance their competitive advantage with a better understanding of the constructs that influence consumers' behavioural intention. We adopted reasonable fit and explanatory power to evaluate these models and to determine which version was best [6]. The findings of the study show that the decomposed theory of planned behaviour model 2 better predicts the users' intention to use e-readers than do other models (although there is no big difference in the goodness of fitness index of the DTPB model 1 and 2). The R2 for each dependent construct is used to assess predictive power. The decomposed TPB model-2 has explanatory power for behavioural intention, attitude, and subjective norm and perceived behavioural control.

Based on the findings, we examined the following ten hypotheses: from H1 to H8, H10, and H12 (attitude → behavioural intention; subjective norm → behavioural intention; perceived behavioural control → behavioural intention; behavioural intention → actual usage; relative advantage → attitude; complexity → attitude; compatibility → attitude; normative influence → subjective norm; utilitarian product → perceived behavioural control). Two hypotheses, H9 and H11, (self-efficacy → perceived behavioural control and hedonic product → perceived behavioural control) were not supported. Our results are in agreement with the result of a previous study by Koeder et al. [10], apart from two hypotheses, H2 (subjective norm → behavioural intention) and H10 (facilitating condition → perceived behavioural control). In terms of H10, their model has the same component as our DTPB-1. DTPB-1 has the component of facilitating conditions and DTPB-2 has the component of utilitarian products. DTPB-1 is not much different from model-2, which has component facilitating conditions, as also shown by Koeder et al. [10]. Utilitarian products have correlations with facilitating conditions. Therefore, we tried these components individually for good model fit.

In terms of non-agreement on H2 with Koeder et al. [10], their results showed that the subjective norm was negatively correlated with behavioural intention for e-reader consumers in Japan; however, our result is positively correlated. The

special role and importance of society in Japan has been discussed in several sociological publications [41], [42]. In a strong social culture such as Japan, it is reasonable to find that there are positive relations between the subjective norm and the behavioural intention to purchase or use e-readers. This finding is in agreement with other research: Taylor et al. [6], Sanjukta et al. [7], Paul et al. [43], Majali et al. [44], Ozer et al. [45], and Mathieson et al. [40] also found a significant relationship between SN and BI.

In our study, the type of product is an important variable that influences consumer choices to purchase specific products (e-books, audio recordings, etc.). Hedonic products are not a significant predictor for the PBC variable but utilitarian ones are. This result coincides with the findings of Sanjukta et al [7]. In that study, they found that the Internet is used more for utilitarian product shopping. Consumers use e-readers when their main focus is on the functional attributes of products/services. E-readers offer various kinds of utilitarian products/services (e-books, applications and so on) to use or purchase, so this will change users' beliefs regarding the opportunities that are needed to perform a particular behaviour. The absence of this construct represents barriers to usage of e-readers. Hedonic products were not a significant predictor for PBC variable, which is again in agreement with previous research [36] in case of Internet shopping behaviour. The absence of hedonic products/services (games, music and so on) may not, per se, encourage users' intention behaviour.

Here, we examined three adoption models: the TRA, TPB, and DTPB (4 different models) with extended antecedents (hedonic and utilitarian products). We chose DTPB-2 as the best model. Utilitarian products are a newly founded component of the model that can improve understanding of users' perceived behavioural control regarding e-readers.

VI. CONCLUSION

This study developed a model with the extension of utilitarian and hedonic products for explaining consumers' behavioural intention to use e-readers. The results indicate that the decomposition of beliefs can provide additional insights into consumers' behavioural intention to use e-readers.

To make e-readers more adoptable, their developers should pay attention to a number of important factors. This study, on the basis of empirical results, provides such developers with suggestions in three areas: (1) system design that makes reading more enjoyable (2) improved network facilities and (3) technological support aimed at increasing e-reader acceptance among consumers. With more complex devices, such as Internet-enabled e-readers, it is essential that they are perceived as relatively easy to use and are compatible with their users' current lifestyle [28]. E-readers should be designed in such way that technology is easy to understand and use. User-friendliness is essential to increasing users' acceptance of e-readers. The newly found factor (utilitarian products) that influences Japanese consumer adoption of e-readers transforms users' beliefs regarding the opportunities that are needed in order to perform a behaviour. Therefore, the absence of utilitarian products represents a barrier to greater adoption of e-readers. Users' main focus is on the utilitarian products of e-readers rather than on their hedonic aspects. This finding will be crucial for the current and future development of e-reader devices. This paper will help to provide a better understanding of user perception of e-readers and what roles cost, connectivity, usability, and content play for users. The DTPB provides useful, easily understood, and relevant information for discerning consumer behavioural intention regarding e-reader adoption.

VII. LIMITATIONS AND FURTHURE RESEARCH

Our study was limited to student sample data. Furthermore, all respondents were from the same university. Future studies could collect data from people of multiple age groups and occupations, and/or from multiple universities, noting that there are correlations between utilitarian products and facilitating conditions. Further investigation into these interrelationships may help to better understand consumer behaviour towards e-readers. The explanatory power of PBC was low and further research is necessary.

REFERENCES

[1] http://en.wikipedia.org/wiki/E-book_reader<06-09-2011>

[2] http://blogs.informatandm.com/1463/press-release-mobile-broadband-e-reader-sales-to-peak-at-14-million-units-in-2013/<06-09.2012>

[3] Voyager (2008).Japanese eBook market. Voyager online.Retrieved from: epub-revision.googlecode.com/…/Voyager%20proposal%20Appendix%202.pdf<06.09.2012>

[4] Fishbein, M. and Ajzen, I. (1975)«Belief, Attitudes, Intention and Behavior: An Introduction to Theory and Research», Reading, MA: Addison-Wesley.

[5] Ajzen, I. (1991), "The theory of planned behavior", Organizational Behavior and Human Decision Processes, Vol. 50, pp. 197-211.

[6] Taylor, S. and Todd, P. (1995), "Decomposition and crossover effects in the theory of planned behavior: a study of consumer adoption intentions", International Journal of Research in Marketing, Vol. 12, pp. 137-55.

[7] Sanjukta P.,Jana H., and Ge X.,2011: Explaining consumers' channel-switching behavior using the theory of planned behavior, Journal of retailing and consumer services 18(2011) 311-321.

[8] Yano Research Institute, 2010: http://www.yanoresearch.com/press/pdf/707.pdf<06.09.2012>

[9] http://www.teleread.com/paul-biba/waiting-for-a-push-the-japanese-ebook-market-in-2011-by-robin-birtle/print/<06.09.2012>

[10] Koeder Marco Josef, Mohammed Upal, and Sugai Philip (2011): Study of consumer attitudes towards connected reader devices in Japan based on the decomposed Theory of Planned Behavior Study of consumer attitudes towards connected reader devices in Japan based on the decomposed Theory of Planned Behavior, Economics & management series EMS-2011-10, 2011-05

[11] Rao, S. and Troshani, I. (2007), "A conceptual framework and propositions for the acceptance of mobile services", Journal of Theoretical and Applied Electronic Commerce Research, Vol. 2 No. 2, pp. 61-73.

[12] Davis, F.D., Bagozzi, R.P., and Warshaw, P.R. (1989), User

 ISBN: 978-1-61208-229-5

Acceptance of Computer Technology: A Comparison of Tow Theoretical Models, Management Science, Vol.35, No.8, PP. 982-1003.

[13] Rogers, E.M. (1983), Diffusion of Innovations, Free Press, New York, NY.

[14] Venkatesh,V., Morris,M.G.,Davis,G.B., and Davis,F.D.(2003).User acceptance model:for longitudinal field studies,management science 46(2),pp.186-204.

[15] MA Jian, 2011: A study on the acceptance and use of handheld E-book readers,Applied mechanics and materials Vols.50-51,pp 209-213.

[16] Sungjoon Lee (2012): An integrated adoption model for e-books in a mobile environment: Evidence from South Korea.telemat.Informat.82012),doi:10.1016/j.tele.2012.01.006.

[17] jaemin J., Sylvia Chan-Olmsted, Bellnine park, and Youngji k.,(2011): factors affecting e-book reader awareness,interest and intention to use.New media & society 1482) 204-224.

[18] malathi Letchumanan and Rohani tarmizi(2011):Assessing the intention to use e-book among engineering undergraduates in University Putra malaysia.Library Hi tech,Vol.29, Iss:3,pp.512-528.

[19] randry Brown (2011): professor acceptance and use of E-readers and E-books in the classroom.The 2011 Las Vegas International Academia Conference.pp 793-796.

[20] Shin-Chun Chou,J ay Stu, and Yutting Lin(20):determinants of E-book readers adoption and continution:A comparison of pre-adoption and post-adoption beliefs. 2010 5th International Conference on Computer Sciences and Convergence Information Technology (ICCIT), **pp.** 853 – 856.

[21] Bram Pynoo, Pieter Devolder, Jo Tandeur, Johan van Braak,Wouter Duyck, and Philippe Duyck (2011): Predicting secondary school teacher's acceptance and use of a digital learning environment:A cross-sectional study.Computers in human behavior 27(2011).pp.568-575.

[22] Jung-Yu lai and Chih-Yen Chang (2012): user attitudes toward dedicated e-book readers for reading: the effects of convenience,compability and media richness. Online Information review,Vol.35.Iss:4.pp. 558-580.

[23] Puschel, J. and Mazzon, J, A. (2010), Mobile banking: proposition of an integrated adoption intention framework. International Journal of Bank Marketing, Vol. 28, No. 5. pp 389-409.

[24] Luarn, P. and Lin, H. (2005), "Toward an understanding of the behavioural intention to use mobile banking", Computers in Human Behaviour, Vol. 21, pp. 873-91.

[25] Pedersen, P.E. (2005), "Adoption of mobile internet services: an exploratory study of mobile commerce early adopters", Journal of Organizational Computing, Vol. 15 No. 2, pp. 203-22.

[26] Laukkanen, T. and Cruz, P. (2009), "Comparing consumer resistance to mobile banking in Finland and Portugal", in Filipe, J. and Obaidat, M.S. (Eds), e-Business and Telecommunications, Springer, Berlin, pp. 89-98.

[27] Riivari,J. (2005),"Mobile banking a powerful new marketing and CRM tool for financial service companies all over Europe",Journal of Financial Service marketing,Vol.10,No.1,pp.11-20.

[28] Suoranta, M. and Mattila, M. (2004), "Mobile banking and consumer behaviour: new insights into the diffusion pattern", Journal of Financial Services Marketing, Vol. 8 No. 4, pp. 354-66.

[29] Malhotra and K. Naresh : (2004), Marketing Research, 4th edition, Pearson Education Education.

[30] Nunnally J (1978). Psychometric Theory. McGraw-Hill, New York.

[31] Fornell,C., Johonson,M.D., Anderson,E.,W., Cha,J., and Bryant,B.E., 1996:The american consumer statisfaction index:nature ,purpose,and findings,Journal of marketing,60(october, pp. 9 7-18).

[32] Hulland 1999: use of partial least squares in strategic management research:A review of four recent studies.Strategic management journal,20(2):195-204.

[33] Green D.E., Morris,T.W., Green J., Cronan,J.E.,Jr., and Guest,J.R.(1995).Purification and properities of the lipoate protein ligase of escheerichia coli.Biochem J.309,853-862.

[34] Menguc,B. and Auh,S. (2006): Creating a firm level dynamic capability through capitalizing on market orientation and innovationness,journal of the academy of marketing science,vol.34,pp.63-73.

[35] Cadogan J.W., Souchon,A.L., and Procter,D.B. (2008): The quality of market oriented behaviors:formative index construction.Journal of business research,61812),1263-1277.

[36] Daire hooper, Joseph C., and Micheal R. M. (2008): Structure equation modeling:guidence for determining model fit,Electronic Journal of Business research Methods.Vol 6,Issue.1,pp.53-60.

[37] Browne, M. W. and Cudeck, R. (1993). Alternative ways of assessing model fit. Newbury Park: Sage Publications.

[38] Rust, R.T., Zahorik, A.J. and Keiningham, T.L. (1995), "Return on quality (ROQ): making service quality financially accountable", Journal of Marketing, Vol. 59 No. 2, pp. 58-70

[39] Hair JF, Anderson RE, Tatham RL, and Black WC (1998). Multivariate data analysis. 5th ed, Upper Saddle River, NJ: Prentice-Hall International,Englewood Chiffs,NJ.

[40] Matheison,K. (1991): predicting user intentions:comparing the technology acceptance model with the theory of planned behavior,Information system research,Vol.2839,173-191.

[41] Benedict, R. (2006). The Chrysanthemum and the Sword, Chicago, IL: Mariner Books / Houghton Mifflin Harcourt Company

[42] Doi, T. (2002). The Anatomy of Dependence. Tokyo: Kodansha International

[43] Paul A. and Lin Chai,(2002): What drives electronic commerce across cultures?A cross-cultural empirical investigation of the theory of planned behavior,J ournal of electronic commerce research,vol.3,No.4,pp.240-252

[44] Malek Al-Majali and Nik Kamariah Nik mat,2010: Journal of Internet Banking & Commerce;July 2010, Vol. 15 Issue 2, Special Section p1 (http://www.arraydev.com/commerce/jibc/)<19-09-2012>

[45] Gokhan Ozer and Emine yilmaz,2011: Comparison of the theory of reasoned action and the theory of planned behavior:An application on accoutants' information technology usage, African Journal of Business management Vol.5(1),pp.50-58.

 ISBN: 978-1-61208-229-5

Towards Enhanced Location-based Services through Real-time Analysis and Mobility Patterns Acquisition

Javier Rubio-Loyola and César Torres-Huitzil
Information Technology Laboratory
CINVESTAV Tamaulipas
Ciudad Victoria, Tamaulipas, México
{jrubio, ctorres}@tamps.cinvestav.mx

Ramón Agüero
Telematics Engineering Group
Universidad de Cantabria
Cantabria, Spain
ramon@tlmat.unican.es

***Abstract*-This paper presents a work in progress towards a middleware platform to support enhanced location-based services through real-time mobility analysis and mobility patterns acquisition. The platform provides services for mobile users and also for mobility analysts. Mobile users are enabled to receive notifications in response to emergency, contingency situations or deviations from mobility patterns. Mobility analysts are enabled to analyse mobility behaviour of users during time scales and territorial scopes as well as to obtain and program mobility patterns and indicators of mobile users and groups of users. Enhanced location-based services are possibly if and only if powerful and efficient capturing, pre- and post-processing of mobile information schemes are adequately implemented and put in place in favour of ubiquitous location service provisioning.**

Keywords-Location-based services; mobile information processing; global position.

I. INTRODUCTION

The study and analysis of mobility and transport aspects have taken on increased importance in recent years, in particular the ones that affect urban sustainability and urban policy [1] [2]. In general terms, current diagnosis of the mobility and transport systems' sustainability base their studies mostly on static models such as the usage of general indicators such as average distances travelled, changes in shifts and changes in the location of productive activities.

To date, there is a lack of mobility systems that deliver real-time services aimed at warning mobile users during a trip with alternative routes in response to emergencies, contingency, and/or congested roads, all in all, considering accurate measurements of mobility patterns. A system like this would be complex, highly dynamic and should be able to deal with a large number of participants, making its applicability difficult for specific contexts such as emergencies, traffic congestion, weather contingency, maintenance of traffic infrastructure, etc., as they should be able to react automatically and they should scale.

In order to assess systems of this kind, there is a need for further analysis of dynamic operational aspects such as the nature of the information available, the type of indicators to use or references to territorial units for ad-hoc applications to particular circumstances [3]. Location Based Services are a set of tools that provide personalized services with help from the user's geographic location or other moving object of interest [3]. These services provide accurate location information through mobile devices such as cell phones, Global Positioning System (GPS) [3] or Radio-frequency identification (RFID) [4]. A location-based service is not necessarily limited to locating and tracking the mobility of a mobile entity. Moreover, these services could be extended towards orientation or navigation services by exploiting the user's location to allow build roads on a map, indicate routes step by step under specific circumstances, determine the distance to travel and display sites that may be of interest to the mobile users [5].

This paper presents our work in progress towards a location-based service platform that provides support for advanced mobility services through real-time analysis of mobility information and generation and usage of mobility patterns. The service platform is envisioned as a platform sensible to dynamic operational aspects such as the nature of available information about routes and mobility, the type of indicator to use, referenced territorial units for ad-hoc applications, and real time and large scale analysis. The targeted system is a location-based service platform whose core functionality is the analysis of mobility information. In particular the platform provides support to help acquiring patterns of mobility that can be used to generate alerts when patterns are not met during a trip of a mobile user. Although the service platform can be thought of being ideal for transport systems, it is possible to use it with services such as security through freight monitoring, assistance for drivers in case of eventual circumstances (e.g., emergency, weather contingencies, etc.), in which notifications are sent automatically to mobile users through the platform.

After this Introduction, Section II presents the conceptual framework of the target platform. Section III presents the preliminary implementation steps, and finally, Section IV concludes the paper.

II. CONCEPTUAL FRAMEWORK

This section describes the framework and challenges of the proposed services platform shown in Figure 1.

A. Conceptual Framework

The platform provides services to mobile users and to mobility analysts, which are described hereafter.

 ISBN: 978-1-61208-229-5

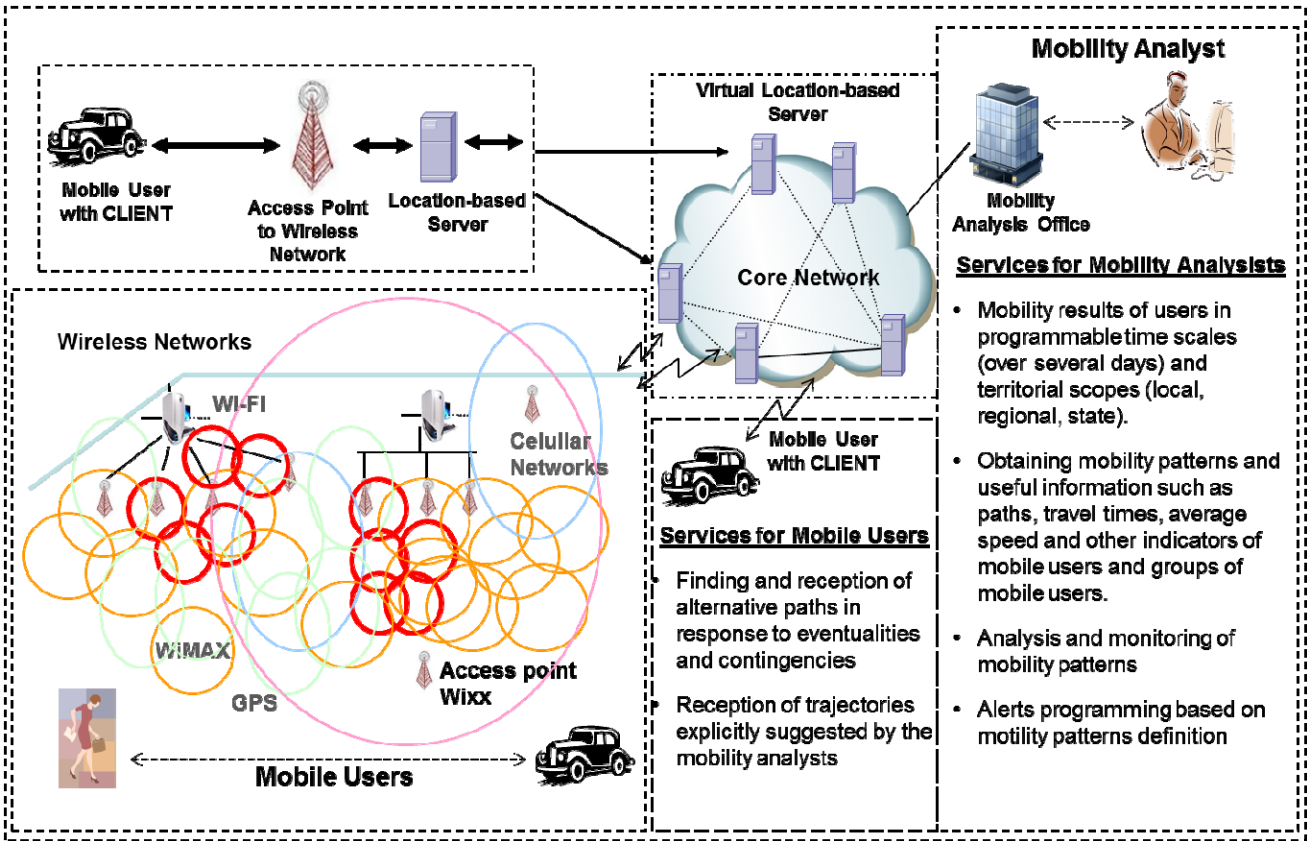


Figure 1. Services supported by the location-based service platform

Figure 2. Example of location service for mobile users

From the mobile user viewpoint, the platform provides support in finding alternative paths in response to eventualities and contingencies, and it also provides support to follow trajectories explicitly suggested by a mobility analyst, who will also be user of the platform. Figure 2 shows an example of the location services for the mobile user, where two alerts are sent to the mobile user as a result of a contingency event, accident, etc. These alerts are sent by the platform as a result of real-time mobility information analysis of other users of the platform.

For the mobility analyst the platform provides support in aspects that include; obtaining results of location and trajectories of various mobile devices during the day and during periods; obtaining mobility results of programmable time scales (over several days) and territorial scopes (local, regional, state); obtaining mobility patterns and useful information such as paths, travel times, average speed and other indicators of mobile users and groups of mobile users. Figure 3 shows examples of services for the mobility analyst. On the left, the trajectory followed by the user in two consecutive days is displayed, where the time traveled, average speed per path segment, and historical information are also available to the analyst (not shown). The center of Fig. 3 shows the area where the user spent more time, time of entry and exit in the two analyzed days. Finally, the right part of Fig. 3 shows the regions of interest, number of visits made by the user, number of mobile users in the region, etc.

The mobile user can navigate to any place and the location data are collected by the platform based on a client-server

 ISBN: 978-1-61208-229-5

design. The platform records the GPS locations of the users' mobile devices. The process of monitoring and subsequent automatic data processing takes place invisibly for both, the mobile users and for the mobility analyst.

B. *Technological Challenges*

For the realization and implementation of the location-based platform described above, it is being necessary to conduct major study on three fundamental aspects, which are briefly described hereafter.

Capturing and pre-processing of mobile information. This aspect deals with the critical nature of monitoring a large number of mobile entities as well as assessing the inter-operation between networks and the need for high bandwidth at critical periods [6]. It also faces the problem of developing algorithms for the generation of location points and the estimation of reliable paths from inaccurate GPS data and missing information due to the unavailability of the GPS signal. Particular problems are the imprecision and uncertainty in GPS data, and the different connotations and semantics of the different types of information from the mobile entities [7].

Information clustering and processing. This aspect faces the problematic of developing algorithms for discovering and predicting mobility patterns under a two-level stream clustering approach: 1) clustering based on time and location; 2) space-based clustering to obtain regions of interest. The central problems of this aspect are: massive information; integration of heterogeneous information from geographical, temporal and population information viewpoints [8] [9]; high computing time and memory requirements required to perform the clustering and processing of location information collected. The latter problem is a motivation to use advanced techniques for optimizing computer resources. Finally, another issue that deserves special attention is the lack of homogeneity in the length of the location data to analyze.

Middleware for the location-based services platform. The purpose of the middleware is to facilitate the development and operation of location-based mobile applications in an environment of technological heterogeneity [10].

The middleware provides filtered information and generates notifications under pre-programmed mobility patterns and/or mobility behavior. The core challenges for the realization of a middleware of this kind are: high data change rates and information updates; need for analysis and interpretation of information at different levels; adaptability and asynchronous interaction; and the aggregation, updates and cancelation of service subscriptions and notifications to mobile users [11].

III. PRELIMINARY IMPLEMENTATION

This section briefly presents the preliminary implementation of the location-based platform described above. In particular, we shortly describe the key aspects of the client-side software architecture that is being currently developed to obtain user's location by using GPS-enabled smartphone devices. Our technological choice is based on the fact that such ubiquitous devices are becoming essential contributors to location-based services as they can provide position information accurately. Smartphones are enabled with a communication channel to send and receive information and therefore, personalized or location-dependent information can be delivered through this channel in order to enhance interaction and deliver high level knowledge. Location data depicting mobility patterns or human behavior can be obtained at large-scale both longitudinally and population-wise.

A. *Client-side implementation*

The preliminary mobile entities implement a client-based architecture to enhance location-based services in smartphone devices. The main goal is to develop a middleware to improve the smartphone software architecture for continuous and efficient services for location data by: a) providing location information at suitable abstraction level, b) collecting and storing meaningful location data, and c) optimizing energy consumption for continuous sensing. Figure 4 shows a simplified block diagram of a device-side model whose main components are briefly described hereafter.

Figure 3. Examples of location service for the mobility analyst

The *Duty cycle adapter* provides the mobile sensing application with the position information. It abstracts the positioning methods and devices (GPS) that can be used to obtain locations and it is in charge of their parameter configuration (adaptive sampling and duty cycling).

The *Orchestration policy module* maximizes the accuracy of monitoring mobility and optimizes location updates according to a sensing policy on diverse smartphone usage scenarios (pedestrian or driving modes) with a given energy budget.

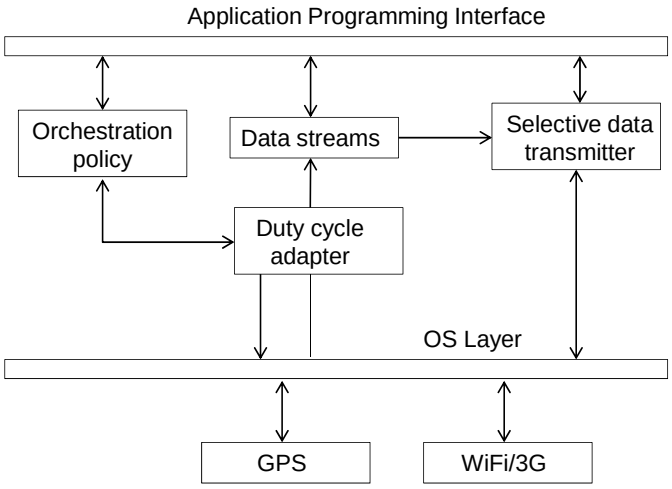


Figure 4. Block diagram of the client software architecture

The *Selective data transmitter* manages local data streams storage and the communication with the server. It provides standard ways to transfer data from/to the clients to/from the server (HTTP [12], HTTPS [12], TCP [12], and UDP [12]) selected according to the applications requirements (e.g., continuous real-time data).

Currently a prototype has been developed built on the software stack on Android-powered smartphones to evaluate the functionality and performance of the client-side architecture.

B. Server-side implementation

Data collected with a mobile device are enough to reveal an interesting pattern on their own. However, when processed through models, and algorithms on series of external and cross-user data sources, simple data can be used to infer complex phenomena about individuals and groups. To make mobility data and location-based services more readily accessible to smartphones, higher level data abstractions are needed at the cost of storage and computation. A preliminary version of a server side architecture is being explored according to the general layered organization shown in Figure 5.

The *communication layer* manages connectivity with the mobile sensing devices. The functionality of this component must match that included in the client-side software architecture. The *data collection and storage* component stores location data in databases. The visualization module shows the information to the mobility analyst. The *data analysis* component analyzes user trajectories at different scales spatially and temporally to automatically extract mobility patterns. This component uses historical data stored in the database to perform inference, correlation, and data analysis tasks to provide a complete view of situations.

The current efforts have been oriented to explore low-level preprocessing techniques for location streams since multiple measurements in the same location do not necessarily yield to the exact same coordinates due to errors and variations in the measurements. For instance, two estimated stay points could have the same semantic meaning, but not necessarily the same exact coordinates. Additionally, geometric and fingerprint based algorithms are being evaluated for the automatic learning of regions of interest since it is a key task to study mobility patterns and human behavior. Such algorithms might be used as the basis for predicting user movements or decision making based on location.

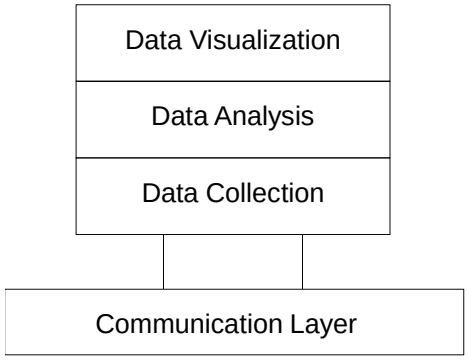


Figure 5. Block diagram of the server-side architecture

IV. STATE OF THE ART

Zheng et al. [8] proposed a method to detect inconsistencies in city planning through long-term analysis of GPS trajectories of taxicabs in urban areas. The method relies on identification and correlation of pairs of regions with salient traffic problems. Even when this method analyses large scale data during long runs, it is somehow preprogrammed to a single pattern of correlation, which is the correlation of pairs of regions with salient traffic problems. The method does not support programmability of mobility patterns (it has only one), and does not deploy location-based services to mobile users at all.

White et al. [9] propose an in-vehicle automatic accident detection and notification system to eliminate the delay between accidence occurrence and first responder dispatch. The system is based on iPhone and Google Android platforms, which can automatically detect traffic accidents using accelerometers and acoustic data, and also the provision of GPS coordinates in case of accident occurrence. This method uses context data to avoid false positives but it does not include characteristics proposed in our platform. To mention some, the

 ISBN: 978-1-61208-229-5

proposed method does not correlate mobile data with other mobile entities or it does not warn mobile users when risky areas for example rainy areas are present in one vehicle's route.

Tzung-Shi et al. [2] propose a method to analyze user movement behavior patterns though standard graph-matching algorithms, which are run over mobile information stored in database systems. Even when the method has been proven to be effective in terms of execution efficiency and scalability, the proposed procedures do not include mobile service deployment of services and the programmability of patterns characteristic proposed in this paper.

Zheng et al. [13] report on a personalized friend and location recommender system for the geographical information systems (GIS) on the Web. Individual interests in (un)visited regions are estimated by involving user's location history and those of other users. The method is based on hierarchical-graph-based similarity measurements to uniformly model individual's location history, and to effectively measure the similarity among users. The proposed system is proven to be effective to find similarity-related metrics like [14] similarity-by-count, cosine similarity, and Pearson similarity measures. However, the method does not provide support to define programmable patterns and exploit them in favor of enhanced ubiquitous services.

All the related work in the literature are mostly intended to analyze acquired mobility information either to deduce mobility patterns, interests, behaviors and so forth. However, the vast majority of works do not allow mobility pattern analysis programmability. To the best of our knowledge, to date there is a lack of systems that exploit mobility pattern information in favor of ubiquitous services provision with energy saving guidelines, and where location-based services can be deployed over mobile clients as response to programmable alerts triggering.

V. CONCLUDING REMARKS

This paper has presented work in progress towards a middleware platform to support enhanced location-based services through real-time mobility analysis and mobility patterns acquisition. The technical challenges for its realization, namely, capturing and pre-processing of mobile information, information clustering and processing and the implementation of a middleware for platform, have been partially addressed. The preliminary implementations of the conceptual framework presented in this paper indicate that the proposal is feasible, and also, have provided some guidelines for its finalization.

ACKNOWLEDGMENT

This work is partially supported by the LACCIR Project No. R1211LAC005, the CONACYT FOMIX project No. TAMPS-2012-C35-185768, and the project TEC2009-14598-C02-02 granted by the MEC Spanish Ministry and partially funded with FEDER funding.

REFERENCES

[1] Ganti, R.K., Fan Ye, and Hui Lei; , "Mobile crowdsensing: current state and future challenges," Communications Magazine, IEEE , vol. 49, no. 11, pp. 32-39, November 2011. doi: 10.1109/MCOM.2011.6069707.

[2] Tzung-Shi Chen, Yen-Ssu Chou, and Tzung-Cheng Chen; , "Mining User Movement Behavior Patterns in a Mobile Service Environment," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on , vol. 42, no. 1, pp. 87-101, Jan. 2012, doi: 10.1109/TSMCA.2011.2159583.

[3] Chon, J., and Hojung Cha; , "LifeMap: A Smartphone-Based Context Provider for Location-Based Services," Pervasive Computing, IEEE , vol. 10, no.2, pp. 58-67, Feb. 2011. doi: 10.1109/MPRV.2011.13.

[4] Henrik Ljøgodt Moen, Thomas Jelle, and Tradløse Trondheim; , "The Potential for Location-Based Services with Wi-Fi RFID Tags in Citywide Wireless Networks". Proceedings of the 4th International Symposium on Wireless Communication Systems, 2007. ISWCS 2007. pp.148-152, 17-19 Oct. 2007 doi: 10.1109/ISWCS.2007.4392319

[5] Shi, W and Liu, Y.; , "Real-time urban traffic monitoring with global positioning system-equipped vehicles," Intelligent Transport Systems, IET , vol. 4, no.2, pp. 113-120, June 2010, doi: 10.1049/iet-its.2009.0053.

[6] Tarkoma, S., and Lagerspetz, E.; , "Arching over the Mobile Computing Chasm: Platforms and Run-times". IEEE Computer 2011 vol. 44, no. 4, pp. 22-28, April 2011, doi: 10.1109/MC.2010.272

[7] Lachapelle, G.; , "Pedestrian navigation with high sensitivity GPS receivers and MEMS". Journal of Personal and Ubiquitous Computing, vol. 11, Issue 6, August 2007, pp. 481-488, doi: 10.1007/s00779-006-0094-3

[8] Zheng, Y., Liu, Y., Yuan, J., and Xie, X.; , "Urban computing with taxicabs". Proceedings of the 13th international conference on Ubiquitous computing; ACM: New York, NY, USA, 2011; UbiComp '11, pp. 89–98

[9] White, J., Thompson, C., Turner, H., Dougherty, B., and Schmidt, D.C.; , "WreckWatch: Automatic Traffic Accident Detection and Notification with Smartphones". ACM Journal Mobile Networks and Applications, vol. 16, No. 3, June 2011 pp. 285-303, doi: 10.1007/s11036-011-0304-8

[10] Bellavista, P., Corradi, A., Montanari, R., and Stefanelli, C.; , "A mobile computing middleware for location- and context-aware internet data services". Journal ACM Transactions on Internet Technology (TOIT), vol. 6, Issue 4, November 2006, pp. 356 - 380, doi: 10.1145/1183463.1183465.

[11] Meier, R., and Cahill, V.; , "On Event-Based Middleware for Location-Aware Mobile Applications," Software En-gineering, IEEE Transactions on, vol. 36, no. 3, pp. 409-430, May-June 2010, doi: 10.1109/TSE.2009.90

[12] J. Kurose and K. Ross; , "Computer Networking. A Top-Down Approach". ISBN 0-13-607967-9 Ed. Addison Wesley 5th edition 2010

[13] Yu Zheng, Lizhu Zhang, Zhengxin Ma, Xing Xie, and Wei-Ying Ma.; , "Recommending friends and locations based on individual location history". ACM Transactions on the Web. Vol 5, No. 1, Article 5 (February 2011), 44 pages. DOI=10.1145/1921591.1921596

[14] Sung-Hyuk Cha; , "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions". International Journal of Mathematical Models and Methods in Applied Sciences. Issue 4, Volume 1, 2007, pp. 300-307

 ISBN: 978-1-61208-229-5

Speech Quality Assessment in Mobile Phones Using a Reduced- complexity Algorithm

Khalid Al-Mashouq
Electrical Engineering Department
King Saudi University
Riyadh, Saudi Arabia
mashouq@ksu.edu.sa

Akram Aburas
Chief Executive Officer
ACES
Riyadh, Saudi Arabia
akram@aces-co.com

Musharraf Maqbool
Communication Head
ACES
Riyadh, Saudi Arabia
khalil@aces-com

***Abstract*—In this paper, we present a reduced-complexity algorithm to assess the quality of speech as perceived by the mobile user. This algorithm utilizes the channel parameters, as measured by the mobile handset, to estimate the speech quality. We used two estimation models; a linear model and a neural network-based model. We compared our estimation with the standard International Telecommunication Union, ITU, objective speech quality measure, perceptual evaluation of speech quality. We found that the linear model can achieve up to eighty four percent correlations with perceptual evaluation of speech quality. Moreover, the neural network-based model can achieve more than ninety percent correlations with perceptual evaluation of speech quality.**

Keywords-Speech Quality Measurement; Perceptual Evaluation of Speech Quality (PESQ); Signal Strength; Bit Error Rate (BER); Frame Erasure Rate (FER); Neural network.

I. INTRODUCTION

Mobile operators are competing to gain customer satisfaction, subsequently reducing the churn rate. Today, voice service is the dominant one among mobile services. Maintaining high speech quality will contribute to better customer satisfaction. The continuous monitoring and assessment of speech quality is essential.

In general, speech quality assessment is performed using subjective or objective methods. A subjective method is based on a group of "good" listeners who can rate the speech signal from 1 (bad) to 5 (excellent). The average score is then taken, which is called mean opinion score, MOS. For obvious reasons, this method cannot be used for the continuous assessment of speech in mobile network.

A subjective method is based on exchanging a "reference" speech segment between a mobile phone to another, preferably, fixed one. The received, possibly noisy, speech segment is then compared with the original "clean" one. International Telecommunication Union, ITU, adopted a standard objective algorithm to process and compares the two speech segments and calculates the quality score, PESQ, or perceptual evaluation of speech quality [1].

Depending on sending a reference speech segment will limit the usability of this method in real-time speech quality assessment. Many researchers investigated other approaches utilizing only the received speech signal, which is called output-based (or non-intrusive) speech quality assessment [2-4]. One approach is to exploit the Markovian structure of speech to detect noise or impairments [2,3]. Another approach is to incorporate some aspects of the human auditory perception mechanism [4].

In general, these approaches are computationally intensive and could affect the processing power of mobile digital signal processor as well as the battery. Moreover, they are generic for any environment and not customized to mobile networks. Distortion in received signal is mainly due to background noise, vocoder imperfection, and/or radio channel noise. Optimization engineer has control only to the later cause. Therefore, our focus is to measure the channel parameters and utilize them to give prediction on the speech quality as affected by the channel impairments.

In this paper, we are collecting a large number of speech samples from a live GSM network using Qvoice (from ASCOM) benchmarking speech tools [5]. Qvoice is used to give PESQ score for all speech samples. In the next section, we describe in details our setup in collecting speech data. Section 3 explains the two different prediction models. They are applied on the collected data and compared with the PESQ score. We outline our conclusions in Section 4.

II. DATA COLLECTION AND PRE-PROCESSING

In this section, we highlight our work which focuses on obtaining speech samples and score them based on the objective speech quality measure PESQ. To facilitate such a testing, we utilized the network benchmarking tool Q-Voice. Random streets from Riyadh city were selected and speech testing was undertaken.

As illustrated in Figure 1, the two major components of the testing system (Q-Voice) are the server and the companion. Pre-recorded speech samples are always stored on the server connected to a PSTN network. The speech samples were carefully selected as "phonetically balanced" to represent normal telephone conversations. The companion who hosts the mobile phones calls the server from the selected streets using the GSM network and once the call is setup, the speech samples are transmitted from the companion to the server. The server upon receipt of these speech samples carries out a comparison and assigns a speech quality score, PESQ, to it.

 ISBN: 978-1-61208-229-5

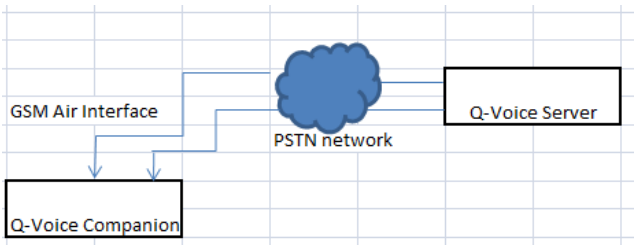


Figure 1. Network Testing Equipment Block Diagram

Upon receipt of these samples at the server end, it is possible to obtain the received speech samples in audio format and verify the degradation level. Speech samples that were transmitted during an active conversation were obtained and parameters associated with each speech sample such as PESQ score, received signal level, RxL, Bit Error Rate, BER, Frame Erasure Rate, FER, and carrier over interference, C/I, were also obtained for the same samples, respectively. Please note that BER is usually mapped under RxQ.

The speech samples obtained from an active conversation were given to listeners with normal hearing. Listeners were trained to familiarize them with the different versions of samples that were received either in excellent or distorted form. Scoring of speech samples by listeners was undertaken in which after hearing each speech sample, the listeners had to grade the speech sample they had listened. This step is needed to randomly verify the machine scoring.

TABLE I. SPEECH SAMPLE GRADING

Score	Classification
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

Table 1 above outlines the grading or PESQ score of the speech samples by the system that was used in obtaining the speech samples and the associated parameters mentioned earlier. Listeners after been trained on some speech samples were then asked to classify the samples as per the table above.

We see it vital to mention here that the algorithm used by the network testing equipment Q-Voice is a reference based algorithm that utilizes the received signal and then extracts associated signal parameters form the speech sample and tries to evaluate to what degree the distortions in the received signal will be audible to the human ear. Every speech sample obtained or used in this exercise constitutes a 5 second voice transmission.

The approach of this paper is systematically aimed at attempting to correlate and successfully link the correlation between the obtained signal parameters and the speech samples graded Qvoice. Table 2 shows and extract of our collected data. Our data contains 759 speech samples each of which is 5 seconds long.

TABLE II. SAMPLE OF DATA OBTAINED FROM QVOICE

RxL	Rx. Qual	FER	C/I	PESQ
-86.56	0.71428	1.28	20	3.9
-83.31	0	1.5	20	3.9
-78.33	0	1	20	3.9
-81.97	0	1.333	20	3.9
-82.76	0.57142	0.9411	20	3.9
-83.10	0	1.2	20	3.9

We move on further to briefly mention about the parameters of importance to us and their impact on a network. RxL is basically an indicator of the coverage been provided by a network operator and is represented as –dBm. RxQ is one signal parameter that is basically a mapping of time averaged bit errors over a scale of 0 to 7 which gives a rough indication on the speech quality. During the analysis, it was found that for every speech sample of 5 second duration numerous RxQ values were obtained, which is quite logical given the duration of the transmission and the fact that the measuring system undertakes the measurement many times. The RxQ values (or the BER values) were averaged out to gain an average for the whole transmitted sample.

Carrier to Interference ratio, C/I, helps in determining the level of interference the subjected signal has undergone. A High C/I will indicate a good signal and yield good communication. Whereas, a low C/I will result in degraded signal quality.

From our pre-processing results, we have also noted that there is a difference on occasions when the human grading differed from the system grading. We therefore, see that there stands a substantial needs for a modified real time network assessment to enable overcome these gaps.

Keeping in view the above parameters and their significance, extensive simulations were carried out on all the samples to feed our estimator. This is shown in greater detail in the following section.

III. PREDICTION MODELS AND RESULTS

We used a liner model to estimate PSEQ score using the following four parameters RxL, RxQ, C/I and FER. They are combined using least square method for optimal weighting. This model is then tested with real data and compared with PSEQ.

The estimated quality score, q, will be

$q = \sum_{i=0}^{4} a_i \; w_i$

where w_i is the weighing factor of the i^{th} parameter

a_1 is RxL

a_2 is RxQ

a_3 is FER

a_4 is C/I

The standard least square solution to this problem is given by [6]:

$\underline{w} = (A^T A)^{-1} A^T \underline{C}$

where; $\underline{w} = (w_1 \; w_2 \; w_3 \; w_4)^T$

 ISBN: 978-1-61208-229-5

and A=($\underline{\alpha}_1$ $\underline{\alpha}_2$ $\underline{\alpha}_3$ $\underline{\alpha}_4$) is the measurements matrix. Here, ($\underline{\alpha}_1$ $\underline{\alpha}_2$ $\underline{\alpha}_3$ $\underline{\alpha}_4$) corresponds to the measurement column vector (of length N) of RxL, RxQ, FER and C/I respectively. Each vector contains N samples corresponding to N speech samples. For, the "PESQ" vector $\underline{C} = (C_1\ C_2\ \ldots..\ C_N)^T$; C_i corresponds to the i^{th} sample of PESQ measurement. We used the whole 759 samples to obtain the optimum linear combination waiting vector $\underline{w}$. The correlation between the predicted PESQ, q, and the actual PESQ is 84%.

The second model is a 2-layer back-propagation neural network [x]. The number of hidden layers is varied between 3 and 30. We used Matlab® neural toolbox for our simulations. Figure 2 shows simulations results when the number of hidden units is 5. The four graphs show the scatter diagram between q and PESQ for training, validation, testing and overall data, respectively. The corresponding correlation coefficients are 0.937, 0.908, 0.942 and 0.932, respectively. It is apparent that the neural model yields good improvement in the prediction ability.

IV. CONCLUSTIONS AND FUTURE WORKS

We have addressed the problem of continuous assessment of speech quality in mobile networks. The purpose is to help operators capture the actual impression of their customers. This should complement the network monitoring and operation centers. We relied on the measured channel parameters to estimate the speech quality of service. We used a linear prediction model, which yielded 84% correlation with the PESQ. A 2-layer neural network is also used, after proper training, to predict the speech quality. The predicted quality score achieved higher correlation, which reaches more than 90%, with the PESQ.

Our approach has reduced complexity compared with Markovain-based ones. This supports its usage within the mobile handset as it would not have significant impact on the processing power and battery life.

The ease of this prediction model can pave the road for several applications. Examples of these applications are

- Customer automated evaluation of the network.
- Quality scores can be relayed to the operator to help in optimizing the network
- Can be used by the operator to give a new tariff procedure or compensation for calls, with bad quality

To make the optimal and robust prediction model, one needs to have much more samples collected over wide spectrum of wireless networks. This should include various geographical locations, different operators and network types.

REFERENCES

[1] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ), An objective method for end to end speech quality assessment of narrowband telephone networks and speech codecs," 2001.

[2] Khalid A. Al-Mashouq, Mohammed S. Al-Shaye. "Output-Based Speech Quality Assessment with Application to CTIMIT Database." Proceedings of the ISCA 17th International Conference Computers and Their Applications, April 4-6, 2002, Canterbury Hotel, San Francisco, California, USA 2002

[3] Chiyi Jin and R. Kubichek, "Vector Quantization techniques for Output-Based Objective Speech Quality," IEEE Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1996.

[4] Kartik Audhkhasi and Arun Kumar, *"Two-scale auditory feature based non-intrusive speech quality evaluation"*, IETE Journal of Research, vol. 56, no. 2, pp. 111-118, March-April 2010.

[5] http://www.ascom.ch/ch-en/tems-symphony-60-datasheet.pdf [Oct 13, 2012]

[6] Lang, Serge, *Linear algebra*, Berlin, New York: Springer-Verlag, ISBN 978-0-387-96412-6. 1987.

.

 ISBN: 978-1-61208-229-5

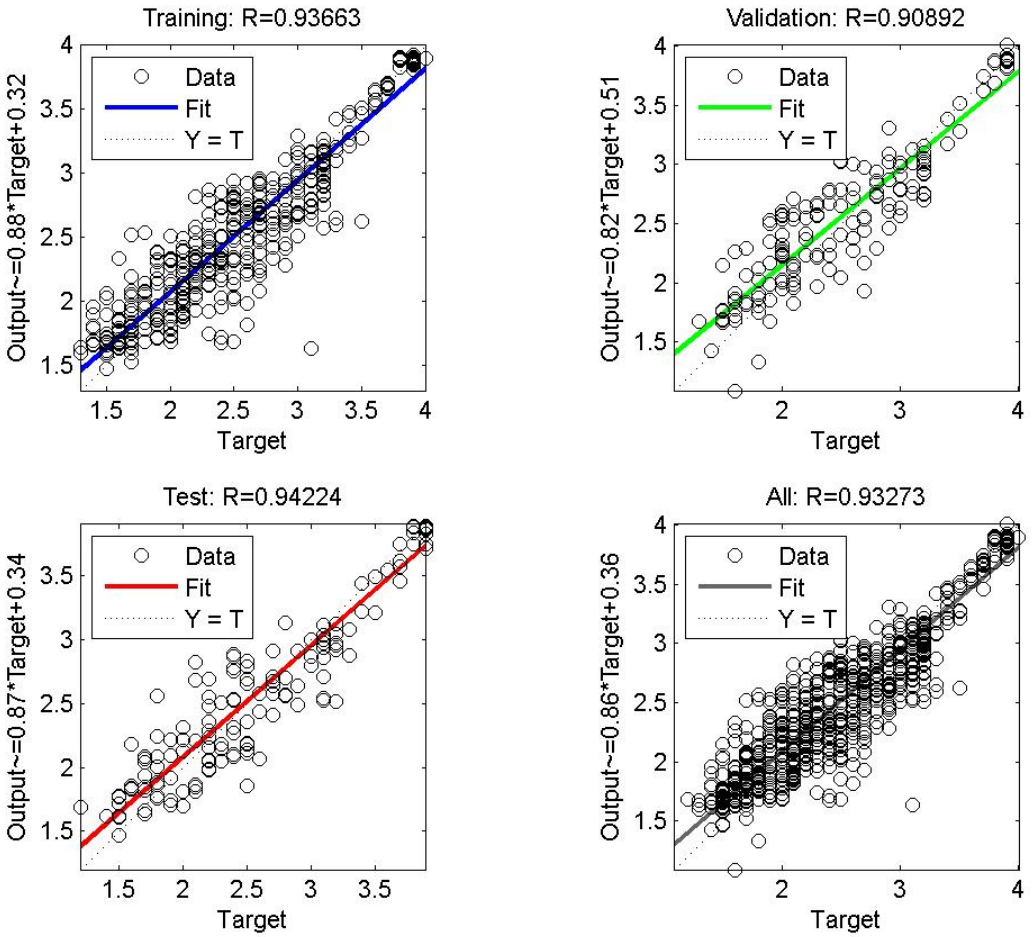


Figure 2. Result of a 2-layer neural network estimator for training (upper left corner), validation (upper right corner), and testing (lower left corner) and over all data (lower right corner). Target stands for PESQ and Output is the estimated PESQ, or q and R is the correlation coeffcient between PESQ and q.

 ISBN: 978-1-61208-229-5

Management of Mobile Objects in an Airport Environment

Gabriel Pestana[1], Augusto Casaca[1], Pedro Reis[2], Sebastian Heuchler[3], Joachim Metter[3]

[1]Inesc-ID/INOV/IST, Lisbon, Portugal
gabriel.pestana, augusto.casaca@inesc-id.pt

[2]ANA-Aeroportos de Portugal, Lisbon, Portugal
pereis@ana.pt

[3]BIJO-DATA GmbH, Heldburg, Germany
sheuchler, jmetter@bijodata.de

***Abstract*—A new approach is proposed to the surveillance of identified Security and Safety occurrences concerning mobile objects in an airport environment, in particular to monitor aircrafts, vehicles and staff at the manoeuvring area for all weather conditions. A middleware platform merges localization information from the different mobile objects in the airport and fuses that information through intelligent algorithms in the platform middleware. The system outputs are shown in an advanced Graphical-User interface, providing a collaborative environment with the relevant information to the airport stakeholders. The outputs can be used by the stakeholders to take decisions on the best way to improve Security and Safety and also on the optimization of airport operational procedures in compliance with existing business rules.**

Keywords-Mobility management; Situation awareness; Airport Safety and Security; Location based services.

I. INTRODUCTION

In the airport environment, about 90% of the accidents and incidents occur during the ground handling services assisting parked aircrafts at the Stand. According to ICAO [1], there is a potential for aviation mishaps to become catastrophes with high casualty numbers. The need for coordination of multiple activities occurring simultaneously requires therefore a continuum control of all ground movements, in particular during taxi operations. However, the current lack of context awareness and controllability is frequently identified as a causal factor for Safety infringements.

Without a unified control infrastructure capable to provide, in real-time, information related to the surveillance of Safety and Security occurrences, airport stakeholders (e.g., Airport Authority, Ground Handlers, Airlines, etc.) have not an objective and reliable view of the overall situation to take well informed decisions in real-time, as needed in various operational domains.

The SECAIR project [2] brings a new approach to the surveillance of identified Security and Safety occurrences for airports, available at an affordable cost, in particular to monitor the mobility of aircrafts, vehicles and staff at the manoeuvring areas in all weather conditions. The SECAIR project combines different location-based technologies to detect the presence of objects (e.g., persons, vehicles), inside the airport terminal or in the apron, at predefined locations. It intends to improve situation awareness for supporting decision makers and task forces in handling with Safety and Security issues. The project relies therefore on the development of an event observer system, capable to identify automatically predefined events and generate alarms in real-time. This means that for each ground movement (e.g., vehicles or any other cooperative moving object), it takes less than one second for the system to determine the new position of the surveyed objects and validate if any of those objects is causing a Safety/Security infringement. A middleware platform provides advanced fusion techniques to determine the localization of objects based on radio based tracking and video based technology. The middleware is part of a larger platform that, on the whole, will manage the mobility of the objects and will enable the accomplishment of an automatic and reliable prediction of hazardous situations.

To test the capabilities of the system, a set of business scenarios addressing airport operational requirements [3] were defined in close collaboration with ANA-Aeroportos de Portugal - the main Portuguese airport's management company, based on the following needs:

- Traceability of vehicles and Ground Support Equipment (GSE) with automatic detection of unauthorised incursions into restricted access areas;
- Tracking and controlling of Handling operations (objects, staff and passengers);
- Surveillance of aircrafts ground movements within the apron area;
- Provision of context awareness about on-going operations at the apron area, triggering Safety and Security alarms with different levels of severity;
- Support the decision making process of the airport stakeholders by providing a reliable view of the

 ISBN: 978-1-61208-229-5

overall situation whenever a Safety or Security event is detected;
- Ensure that each airport stakeholder has access only to data according to its operational needs.

The paper is organized as follows: Section II presents the main software components within the multi-tier architecture designed for the SECAIR system. Section III presents the environment where the project will be deployed together with the operational scenarios defined for testing the system. Finally, conclusions are included in Section IV.

II. SYSTEM ARCHITECTURE

The SECAIR system has a client-server architecture, structured into three tiers, as outlined in Figure 1. At the communication tier, SECAIR will operate with heterogeneous wireless location-based technologies (sensors), each one sending data, in real-time, about the location of the tagged objects. At the application tier, the middleware software component is responsible to collect and process incoming data from the wireless sensors, delivering reliable location data to the Business Logic. This is performed based on a data fusion process that computes positioning data to provide accurate and reliable location data about the surveyed object.

Since the SECAIR system will operate with heterogeneous sensors, prior to the data fusion process, it receives multiple positioning data originated in the mobile objects from the communication tier. In fact, we can have a set of data for one object. But after the data fusion process, we obtain one computed position per object that is reliable. Based on such issues, the SECAIR system will provide a set of innovative capabilities for positioning accuracy and reliability, which is required more and more by value-added location-based services such as Safety and Security applications for airports [4].

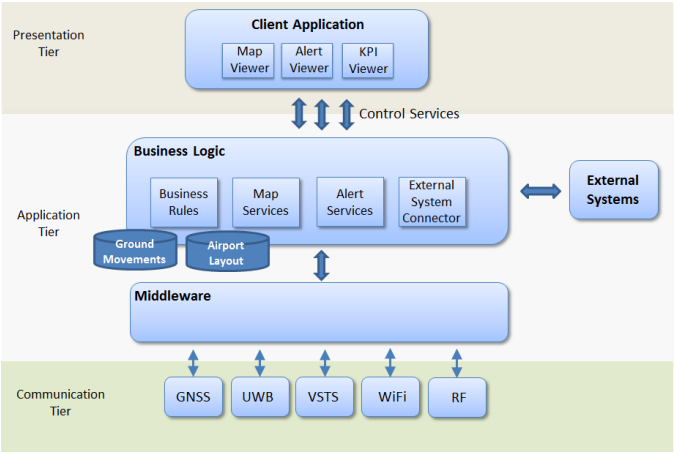


Figure 1. Architectural structure of the SECAIR system.

The Communication tier will operate with the following localisation technologies:
- The Stand-alone Global Navigation Satellite System (GNSS), will be used together with a WiFi communication device to collect and transmit, in each second, the coordinates of the vehicle position;
- The Ultra Wide Band (UWB) system provides immunity to multipath propagation and precision range measurement capability. The IEEE 802.15.4a UWB standard implements precision location measurement. when tracking items close to large metallic objects such as Aircraft, Vehicles, Cargo containers etc.;
- The Video Surveillance and Tracking System (VSTS) consists of multiple video cameras installed at predefined locations so that they fully cover the monitored area with overlapping field of view. The video data collected from the cameras will be processed by the VSTS to detect, track and classify the foreground objects within the monitored environment;
- The Radio Frequency (RF) locating system consists of mobile devices (tags), antenna units (anchors) equipped with smart-antennas and mounted in the area of interest. It measures the position of a mobile device attached to the person or the object to be tracked in the area of interest (see Figure 2).

The Middleware is responsible to continuously provide the calculated position of a specific observed object to the Business Logic. A major concern regarding the project real-time effectiveness and positioning accuracy is reflected on the analysis of the fusion algorithm which follows a multi-particles approach. The particle filter is a technique that implements a recursive Bayesian filtering using the sequential Monte Carlo method, currently one of the most advanced techniques for data fusion. Three concepts for the data fusion process are considered by the Middleware component:
- Quality of positioning-based selection;
- Sequential Kalman Filtering [5] for simple data fusion without map filtering for simple localisation error distributions;
- Sequential Bayesian Filtering (Particles filters) [5] suitable also for the map filtering.

There is no single technology which can provide satisfactory performance in all environments and scenarios; therefore, various localisation technologies have to collaborate in order to deliver a flexible locating system, instead. Sensor data fusion will combine sensory data from different localisation technologies to outperform any individual systems working alone. These are lessons learned from the LocON project [6] in relation to techniques for multi-target, multi-sensor tracking.

At the Application tier, the Business Logic aggregates a set of core software components. For each ground movement, the Application tier takes descriptive data (metadata) to correlate existing business rules with the spatial-context of the airport and if applicable trigger the right event. The integration of all these data enables the system to perform a set of validations based, for instance, on the type of object, its location within the airport, the operational status of the areas where the object is located.

 ISBN: 978-1-61208-229-5

Depending on the business rule being infringed a specific event (i.e., alert message) is triggered to the end-users at the Presentation tier. This is done by creating a subscription offered as a public endpoint by the system.

At the SECAIR system every location based services run in parallel, and is accessed via Network Load Balancing, a clustering technology that enhances the scalability and availability of mission-critical, TCP/IP-based services. Since every service runs in parallel, the number of instances on different machines is unlimited.

The Business Logic sends the requested data, either as a stream of updates (event-based queries) or as a chunk of current state data (instant-queries). The first are triggered on a certain event, e.g., an object moving into an area. The Business Logic can create an event subscription ("tell me about objects moving into a specific area") to be notified on that event ("an object moves into an area") and perform specified actions accordingly ("alert: object moved into restricted area"). This kind of subscription may be triggered very often, or never, depending on how often the event occurs. Contrarily, the result of an instant query is always returned immediately and is not dependent on any event. This kind of query is useful to retrieve the current state of an object. For instance, "give me a list of all objects which are currently in a certain area" or "tell me the current battery status of an object".

For simplicity, the core data handled by the Business Rules software component are represented in Figure 1 by two distinct databases. These data are managed by the Business Logic using the Microsoft SQL Server 2008, a database management system capable to deal with business data and map features, describing the airport cartographic layout, within the same database.

In the SECAIR system, all thematic layers use the World Geodetic System 1984 (WGS84) as the spatial reference system to fulfil the requirements defined in the A-SMGCS manual [2] and to comply to the ED-119 standard [7]. Therefore, horizontal locations are provided as latitude/longitude coordinates. Each layer can be managed as an information set independent of other layers. Since each layer is spatially referenced, they overlay one another and can be combined in a common map display. The user can then interact with the features of each layer by selecting, for instance, a specific stand and manually change its status, or to get information about flights, resources and assigned tasks. It is also possible to verify which road segment is operational and check for traffic circulation rules that apply to a selected road segment (e.g., speed limit and directions of traffic flow). In this case, authorized users can even specify speed limits for each road segment for different visibility condition.

The resulting geo-database consists of vector and attributes features. The vector features represent geometric feature instances that are classified as points, lines, or polygons. Examples include runway thresholds, holding lines, and aircraft stand locations. The vector features can also represent obstacle data elements, which may be represented by points, lines, or polygons.

The ED-119 standard defines the physical dataset requirements that have been followed to develop the airport mapping. These include: geometry and quality, feature rules and descriptive attributes.

The Business Logic also holds a software component that is responsible to handle the interoperability with external systems, for instance, to collect data related with flight schedules, resources and assigned tasks. With such approach, location based data for each observed object can be coherently correlated with metadata from external sources, enabling the surveillance and track of events according to business logic/rules [8]. The Application tier, being responsible to implement airport business logic, seeks grounds for the coexistence and balance between the dual trends of the airport industry: increased demand for air travel and strengthened aviation Safety and Security [9].

For each predefined event (e.g., Safety or Security) detected by the system, a semantic meaningful alert message will be triggered, with the corresponding relevance and severity risk. These functionalities is provided by a geographic information system (GIS) specifically designed to handle with the business logic taking into account the airport spatial context provided by the Map Services. Depending on the nature of the detected event, the Alert Services will interact with the GIS to generate an alarm to be broadcasted to each connected client application. A log record of all events is stored for historical data analysis purpose.

At the Presentation tier, the surveillance capability of the SECAIR system is presented to end-users in three different ways. The Map Viewer represents moving objects as colour coded point features with a timestamp and a set of descriptive data about the resources causing, for instance, a Safety event; this may include data about the aircraft (A/C), vehicle, driver, flight data, airport layout of the area where the event occurred. The Alert Viewer shows the corresponding textual description of the alert messages in terms understandable by the end-user (e.g., for each moving object causing an event, the Alert Viewer at each Client Application will present the alert messages contextualized with business semantic and ordered by severity level). All alert messages have a start and end time, plus a set of additional descriptive data related to each event. The KPI Viewer presents in a spatial dashboard, the values of key performance indicators (KPI) describing how the business is performing.

The correlations between KPI are mapped in a dendrogram structure. Each individual KPI are arranged along the bottom of the dendrogram and referred to as leaf nodes. KPI clusters are formed by joining individual KPI or existing KPI clusters with the join point referred to as a node, forming a node-link tree diagram. Each node of the tree carries some information needed for efficient plotting or cutting as attributes, of which only members, height and leaf for leaves are compulsory:

- Members, total number of leaves in the branch;
- Height, numeric non-negative height at which the node is plotted.

 ISBN: 978-1-61208-229-5

The hierarchical structure of the dendrogram is represented by the KPI Viewer using the Squarified Treemap algorithm [10]. The Treemap technique provides an area-based visualization where the size of each rectangle represents the relevance of the KPI and the color indicates how the value of the metric is evolving. The Treemap technique is indeed very effective in showing the attributes of the dendrogram nodes using size and colour coding.

The Squarified Treemap algorithm avoids the generation of thin rectangles, improving the representation of the dendrogram structure in a space-constrained layout. It also enables end-users to compare nodes and sub-trees even at varying depth in the dendrogram, and help them spot patterns and exceptions.

III. CASE STUDY

In order to validate the SECAIR system, a system prototype for a pilot test will be installed at Airport of Faro (AFR), Portugal. AFR is one of the Portuguese transport infrastructures included in the Priority Project 8 - Multimodal axis Portugal/Spain-rest of Europe, 2009-PT-08006-E. Acting mainly as a gateway for tourists who predominantly visit the Algarve region and the Spanish region of Huelva, AFR operates mostly with low cost carriers in a seasonal basis with its peak at Christmas and Summer time.

The implementation comprises the system deployment, the interfaces to heterogeneous localization technologies and a set of client applications with a geographical interface for airport stakeholders to benefit from the control services provided by the system. For field tests, ANA-Aeroportos will provide airport vehicles together with a wireless network covering all airport operational areas. The manoeuvring area of AFR is already equipped with an infrastructure of Wi-Fi Access Points (AP) forming the wireless network that will support the data communication.

Figure 2 outlines the areas selected for the specified scenarios and to be used for the site test, namely indoor environment, adjacent indoor-outdoor transition area (to demonstrate ability to track targets moving from indoor to outdoor and back) and the apron area adjacent to stands 14 and 16. Indoor environment scenarios (Boarding Gates 01 and 02), include: zone intrusion detection, target tracking and left behind luggage. The outdoor environment (130x130m of apron comprising aircraft stands 14 and 16), include additionally the following situations: Aircraft Stand Area Vehicle Surveillance (vehicle tracking, obstacle detection), Incursion/Collision Avoidance and Aircraft Ground Movement Tracking.

As presented in Table 1, indoor scenarios reflect operational procedures related to:

- Traceability of a person at the boarding gate area;
- Localization capability of the SECAIR system in the transition area from passenger terminal into restricted access areas (outdoor);
- Localization obtained by fusion of data obtained from the following technologies: VSTS and RF.

The outdoor scenarios reflect operational procedures related to:

- Traceability of vehicle and driver at the Apron area;
- Automatic detection of drivers without driving permission / not logged (RFID Reader)
- Localization obtained by fusion of data obtained from the following technologies: VSTS, GNSS and UWB;

The airport layout is represented (at the Map Viewer) as a collection of overlapped themes, each one representing a specific operational area within the airport environment. These themes form the background context over which the observed objects are represented as point features. All thematic layers are provided by the airport authority in a standard format as shape files.

Whenever an object causes a hazardous event, the Business Logic uses the metadata provided by each theme for location-awareness purposes and generation of the proper alert message to be sent to the Presentation tier.

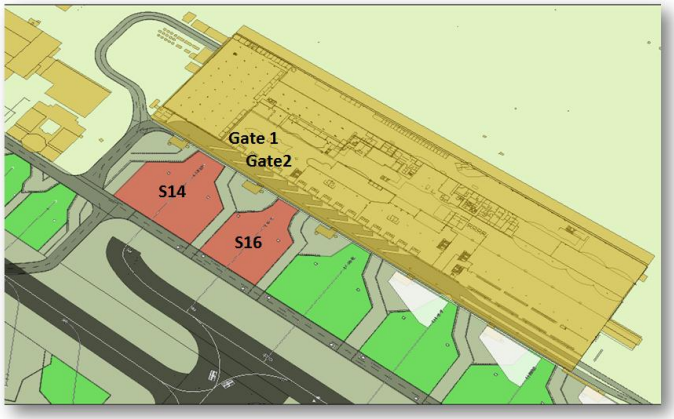


Figure 2. Airport layout of the manoeuvring areas selected for the specified scenarios.

The GIS engine will have to deal with real-time (i.e., each second) requirements for surveillance of all moving objects, computing simultaneously dynamic changes to the spatial context derived from daily airport business activities. Additional metadata (e.g., speed, logged driver, vehicle category, task, etc.) about each surveyed object, presented as labels, provide airport stakeholders with all the required information to analyse a specific event or to remotely monitor/coordinate on-going operations. Moreover, the ability of the GIS to graphically present information from heterogeneous mobile sources enables the system to be operated even by non-skilled experts. This is particularly relevant to monitor the stands' supporting areas where vehicles and GSE are allowed to park in specific technical supporting areas, before an A/C entering the assigned stand are also included.

Some preliminary tests were performed to collect field data to be used for testing during the development phase. The tests included the collection of a huge amount of data for both Indoor and Outdoor environments, involving personnel

 ISBN: 978-1-61208-229-5

and some vehicles and a parked A/C at stand 14, for embarking and disembarking procedures. Within the scope of the SECAIR project, the scenarios presented in Table 1 address Safety/Security issues.

TABLE 1. LIST OF OPERATIONAL SCENARIOS FOR THE SECAIR PROJECT.

Num.	Scenario Name
Outdoor Safety (OSA)	
OSA.01	Surveillance of vehicle movements within a stand area
OSA.02	Collision avoidance support service
OSA.03	Aircraft ground movement tracking
OSA.04	Obstruction of an operational stand area
Outdoor Security (OSE)	
OSE.01	Detection of zone intrusion by unauthorized vehicle
OSE.02	Personnel tracking at the apron area
Indoor Safety (ISA)	
ISA.01	Working zone intrusion by unauthorized person
Indoor Security (ISE)	
ISE.01	Left luggage detection
ISE.02	Indoor-outdoor personnel tracking

The description of each operational scenario follows a template emphasizing relevant issues from airport stakeholders' point of view. Besides a unique identifier with a semantic meaning for each scenario the template also covers the following items:

- Name of the scenario, pointing out concerns from the perspective of airport stakeholders;
- Classification of the scenario addressing environmental influences (indoor/outdoor) and type of events (Safety/Security);
- Technical constraints and a list of key indicators captured by the scenario to measure its impact or relevance;
- List of actions to be performed by each intervenient actor to test the specified scenario;
- Identification of the scenario expected results. This will define the behaviour of the SECAIR system.

Some vehicles will be equipped with an onboard unit, a touch screen display and a radiofrequency reader. A radiofrequency card, with data about the driver involved on the site tests, will be available to automatically identify the driver at the login procedure. At least two client applications (i.e., situation rooms) are deployed, one situation room corresponding to the Control Centre for the airport operator and a second situation room for another airport stakeholder (e.g., Ground Handler).

IV. CONCLUSIONS AND FUTURE WORKS

Within the SECAIR project, one of the technical requirements is that all location–based technologies are coherently integrated using advanced data-fusion techniques in order to reduce installation costs and to address multipath effects reduction. The main objective is to develop new context aware services based on an innovative solution integrating high-performance RF tracking combined with optical recognition technologies and mobility management in a middleware platform.

SECAIR is being designed as a heterogeneous sensor fusion system architecture, covering the surveillance of non-cooperative resources and functionalities for continuous control of all ground movements within the apron area. A special attention is being given to the environment of the system, in particular to information flowing from and to the system. The properties of the system components, as well as the relationships between them, are core elements for the analysis and design of the SECAIR architecture. The project takes lessons learned from previous projects in relation to techniques for multi target, multi sensor tracking, responding to very important Security and Safety issues in airport environments

The software components of the SECAIR system are being tested and the results analysed to confirm the ability to improve positioning accuracy and reliability, reduce the likelihood of false alarms and get feedback from airport stakeholders for improving the system-of-interest and to study its feasibility with real data, as required by value-added location-based services.

A field test is planned to take place at Faro airport during the last quarter of 2013, with a full evaluation of the results to be done at the beginning of 2014. The research work in the project addresses data integration from the video system with data from the radio based systems. Improvements to accuracy and scalability of the location based services, provided by the system, will continue until middle of 2013 and from then on the necessary implementations will be completed for the field test.

ACKNOWLEDGEMENTS

The SECAIR (ref. E6030) is an R&D project, partially funded under the EUROSTARS program, which started in September 2011. It is also acknowledged the funding from FCT - Fundação para a Ciência e a Tecnologia through the PIDDAC program.

REFERENCES

[1] ICAO: Safety Management Manual (SMM), 2nd ed. Doc 9859 AN/474, ISBN 978-92-9231-295-4, 2009.

[2] Eurocontrol: Operational Concept and Requirements for A-SMGCS Implementation Level 2. Ed. 2.1, 2010.

[3] SECAIR Deliverable 1.1: Definition of requirements and operational scenarios. Technical report presented to the Eurostars, Dec 2011.

[4] M. Ayres Jr. et al.: Safety Management Systems for Airports: Guidebook. Volume 2, ACRP REPORT 1, ISBN 978-0-309-11798-2, 2009.

[5] LOCON Deliverable 5.1: Concept of High Level Sensor Fusion. Technical report presented to the Eurostars, Jun 2009.

[6] G. Pestana, N. Duarte, P. Catelas, and J. Metter, Technical Document: LocON Client GUI Specifications, available on http://www.locon.org, Oct 2010.

[7] EUROCAE: ED-119B : Interchange Standards For Terrain, Obstacle, And Aerodrome Mapping Data, 2011.

[8] N. Subbotin: Development of an Airport Ground Vehicle Runway Incursion Warning System. DOT/FAA/AR-11/26, 2011.

[9] G. Pestana, N. Duarte, I. Rebelo, and S. Couronné: Adressing stakeholders coordination for airport efficiency and decision-support requirements. Journal of Aerospace Operations (JAO11), 2011.

[10] G. Chintalapani, C. Plaisant, and B. Shneiderman: Extending the Utility of Treemaps with Flexible Hierarchy, in Proc. of Int. Conf. on Information Visualisation, London, 2004.

 ISBN: 978-1-61208-229-5

A Novel Framework for Personalized and Context-aware Indoor Navigation Systems

Attila Török and Tamás Helfenbein
Institute for Applied Telecommunication Technologies (BAY-IKTI)
Bay Zoltán Nonprofit Ltd. for Applied Research
Budapest, Hungary
Email: {attila.torok, tamas.helfenbein}@bayzoltan.hu

***Abstract*—The recent indoor localization techniques use inertial sensors for position estimations in order to obtain a certain degree of freedom from RF solutions. Unfortunately, this dependency cannot be completely eliminated due to the cumulative errors introduced in the localization process; thus, RF or visual reference points are still necessary. In this paper we propose a novel approach for architectural design of indoor localization and navigation services by introducing a context-aware and extendable system framework. We exploit the ability to recognize certain human motion patterns and by using a scenario specific navigation language, for guiding the localization and position refinement process, we will be able to control the navigation on a much finer level. Therefore, in our system the reference points become needless or for scenarios with topological black holes at least the refinement process is automated, the user can be omitted from finding these points.**

Index Terms*—indoor navigation; context-awareness; location based services; pedestrian localization.

I. Introduction

With the advent of smart phones Location Based Applications (LBA) [1] witness an ever increasing popularity. While commercial services mainly focus on outdoor use cases indoor LBAs suffer a relative backlog, although at first sight all the necessary building blocks [2] are available. Besides the lack of common standards an even more stressful reason can be attributed to the scenario specific nature and sensibility to infrastructural changes of indoor localization techniques. Existing solutions require special effort to build detailed RF maps or propagation models and these pre-deployment steps must be repeated in case of variations (topology, transmission power) in system configuration. Therefore, the research community has started to focus on indoor positioning techniques where pre-deployment efforts are not necessary [3] or where the localization is based on minimal infrastructure [4]. To obtain a certain degree of freedom these systems leverage the technological advancements in mobile devices, such as the inclusion of accelerometer, compass or magnetometer sensors [5] [6]. However, in the current proposals the dependence from some kind of reference points (RF, visual information or acoustic beacons) cannot be completely eliminated, since due to the nature of the inertial sensors by distancing from the last known reference position cumulative errors will be introduced in the location estimation process.

In recent indoor navigation systems, besides the afore mentioned inertial sensor fusion (called dead-reckoning (DR)) techniques, camera phones can also support localization and navigation. These approaches [7] [8] [9] use well placed visual markers (e.g., 'YOU-ARE-HERE' (YAH) maps, QR codes) in order to provide reference points for DR drift cancellation. To further improve navigation experience besides the traditional map based solutions augmented reality (AR) interfaces are also applied. The traditional AR interfaces usually require continuous localization of the user, while in newer ones constraint diminution is achieved by using sparse localization techniques. Unfortunately, these solutions still require reference points [7] [9], constant user interaction/supervision [10] or an occasional manual reset of the accumulated location error [8].

A common problem with current indoor LBA solutions is the moderate effort dedicated to consider and deeper explore the dimensions of contextual relationships in the localization and navigation process, consequently the different requirements arose from personalization (user preferences/capabilities), scenario peculiarities (topological/service types) and their relationship to positioning/route guidance is not integrally handled. Different users will have different capabilities and requirements regarding the navigation procedure. For example, active participation in the process (navigation or interaction through AR interfaces) should be avoided, since it can distract the user, causing confusion/accidents. Instead, a proper voice guidance based navigation service shall be used. Considering scenarios, in certain premises the placement of any kind of reference points is beyond possibility due to legal-, investment issues, or their usability is just simply questionable. Also different levels of quality of service will be required for positioning and route guidance in miscellaneous scenarios.

In this paper, we propose a novel approach for architectural design of indoor localization and navigation services, aimed to provide an extendable framework for personalized and context-aware indoor LBAs. Our goal is to provide a navigation system, which requires no RF infrastructure and where the user interaction for finding the reference points is minimized or at least is automatically triggered by the scenario, the navigation process itself. This requires the employing of human movement behavior analysis, the introduction of a special navigation language, which controls the localization and position refinement process, and the design of a novel architectural framework to empower and piece together the building blocks of the system.

 ISBN: 978-1-61208-229-5

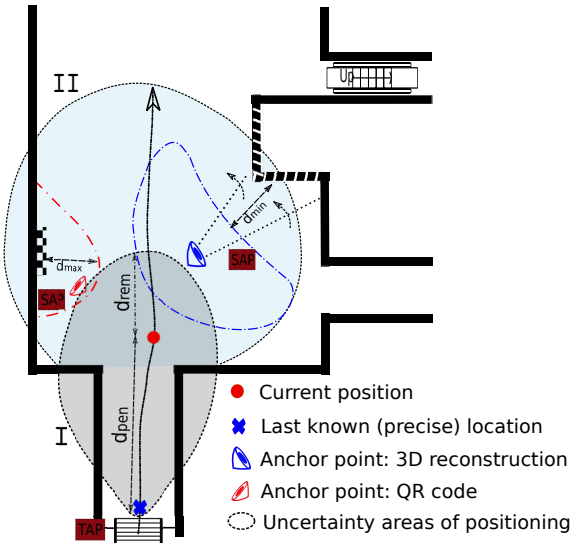


Fig. 1. Problem definition of indoor positioning

The paper is organized as follows. We present the identified indoor localization problems in Section 2, followed by the description of the system architecture. Section 4 presents the position refinement using cameras and we conclude the paper in Section 5.

II. Problem Statement

Navigation systems based only on inertial sensors are corrupted by cumulative errors. As this error grows, the positioning inaccuracy can cause local disorder of the navigation service. Refinement of the position estimations is done by fusing the results with absolute position measurements derived from the reference points of an additional localization system.

Since in our system we want to keep away from using reference points, which require mounting of special infrastructures (e.g., WiFi access points) we have to look for new ways to provide absolute positions for localization error cancellation. Considering, for example a floor map of a subway system we can identify certain building blocks like underpass areas, halls and platforms usually connected with tunnels, stairs and escalators. By traversing trough such places we will generate specific behavior patterns, such as walking, climbing stairs, making turns, taking elevators/escalators, etc. These actions can be considered as special events and by correlating the recognized events with the estimated movement pattern and the topology of the floor plan we will be able to refine the position estimation of the users. These topology specific points used for localization error correction we call Topological Anchor Points (TAPs). Therefore, we extend the functionality of inertial sensors to detect human motion related patterns by using real-time feature detection algorithms at the mobile side. In the current localization systems the accelerometer sensors are mainly used as a digital step counters [4] [5], the recognition of human activities is typically used in Assisted Living applications to detect daily activities of people (e.g., [11]) in health-care services.

Despite the introduction of this concept in certain scenarios there will be no specific topological points, which can generate particular, well interpretable events or the distance between two consecutive TAPs will be too long. In such cases (called topological black holes) the growth of the localization error (uncertainty area nr. II on Figure 1) cannot be suppressed efficiently, this leads to inaccuracy of positioning and to possible navigation problems (missing the second exit with the escalator). To cope with similar situations we also allow to create Soft Anchor Points (SAPs), which are not inherently topology specific and where the special events triggering error cancellation are derived by using additional techniques (e.g., by identifying visual markers like QR codes or 3D positioning via cameras). These SAPs will be defined by the users/service administrators, their usage requires some sort of user interaction, similarly to the visual reference points used in [7] [8] [9]. However, in our system the placement of SAPs is not hard-coded in the system, we do not expect the existence of fixed visual reference points (e.g., YAH maps), since in many cases these are not optimally placed, considering the navigation service needs. The possibility to dynamically place the SAPs can also facilitate the introduction of new location refinement methods, such as computer vision techniques. Since the SAPs are placed to unknown locations we have to notify somehow the users for triggering the refinement process.

To analyze the problem of SAP placement and refinement triggering we present the scenario of Figure 1, where we can observe a certain correlation between the favored placement of SAPs and the movement pattern of the user. As we can see, the last known precise user location is triggered by a TAP (stairs) and from this point the positioning error grows (uncertainty area) as the user enters into the hall. Despite the increasing localization error, until the user overpasses the uncertainty area nr. I (by traversing through the tunnel and entering the hall), there is no reason to trigger the refinement process, since no useful information can be provided for the navigation process. The growth of the localization error can be handled by DR and a context specific mechanism, which will keep the navigator on track, not letting to assume any unnatural events (e.g., crossing the tunnel walls), since is physically impossible to act differently due to the scenario's constraints. As the uncertainty area grows beyond a certain, well defined level, and the user is possibly situated in a more complex scenario (hall/underpass with many exits) the refinement process has to be triggered. This point of action will be calculated based on the floor plan, the nature of the inertial localization algorithms (estimated positioning errors) and the user preferences (e.g., do not use camera). Choosing a specific method for SAP creation will also affect the quality of the navigation process. For example, by using computer vision the area from where correct localization can be effectuated is larger than using QR codes, and it is also easier for the user to perceive and find the right spot (zones around SAPs). The context specific evaluation will also let us to provide an optimal placement for the SAPs, considering the floor plan, the navigation service requirements and the existence of scenario specific TAPs.

 ISBN: 978-1-61208-229-5

III. System Architecture

In order to preserve the flexibility of the framework, we propose the system architecture presented on Figure 2. One key design concept is the separation between the service specific and sensor related data plans. Service specific data (e.g., map tiles, route queries/responses, navigation instructions) are exchanged through the communication channels between the mobile and the respective service. This channel is also used to provide a navigation specific script for the mobile application, which we call Navigation Markup Language (NML).

To assure extendibility, all the information involved in the localization process is considered as sensor data and is exchanged through a Data Gathering Server (DGS); thus, we will be able to provide location related information for other services, too. The output of some sensors (usually accelerometer and magnetometer data) is processed at the mobile side by defining proper signal processing and classification rule sets, while other sensor's data (e.g., GPS) is collected as a result of a simpler query. The first kind of sensors (called Virtual Sensors (ViSe)) will let us to specify the feature detection algorithms used for human motion recognition, giving the ability to derive information for the localization algorithms (e.g., distance calculation based on step counter, DR) and also for the error correction algorithm using TAPs (e.g., by producing events in case of stair detection). The sensor data involved in localization and navigation is shared through the ViSe Routing Nodes, using a publish/subscribe communication graph. The capabilities, the ViSe defined on a specific mobile node, are published and shared through a commonly accessible Knowledge Base (KB), from where all the newly introduced services can acquire information (ViSe discovery). ViSe data and control information are separated using content based routing in DGS (by default the KB gets only control data), in order to facilitate the scalability and extendibility of the system. Thus, the transport technology used between the mobile and the service (e.g., XMPP) is detached from the service discovery, the method of accessing the KB (web service). Inter-service communication and eventing is also better supported, since the service advertisements can be channeled into the KB. The control of mobile sensors, related to the requirements of a specific service, is done by using the Service Control Nodes, whose membership has to be managed by the service or the KB itself.

In our system the Navigation Markup Language (NML) is used to control the localization and to provide context specific information besides the usual navigation related information (floor plan, navigation instructions). The NML is generated during the route planning phase, by analyzing the scenario itself (topology of the indoor environment, user preferences, etc). From the topology graph derived from the floor plan and the planned route the affected TAPs will be determined, acting as error cancellation points. The ViSes on the mobile related to the specified TAP recognition (providers of the respective movement patterns) will be asked for event reporting in form of queries. Finding the corresponding SAPs along a planned

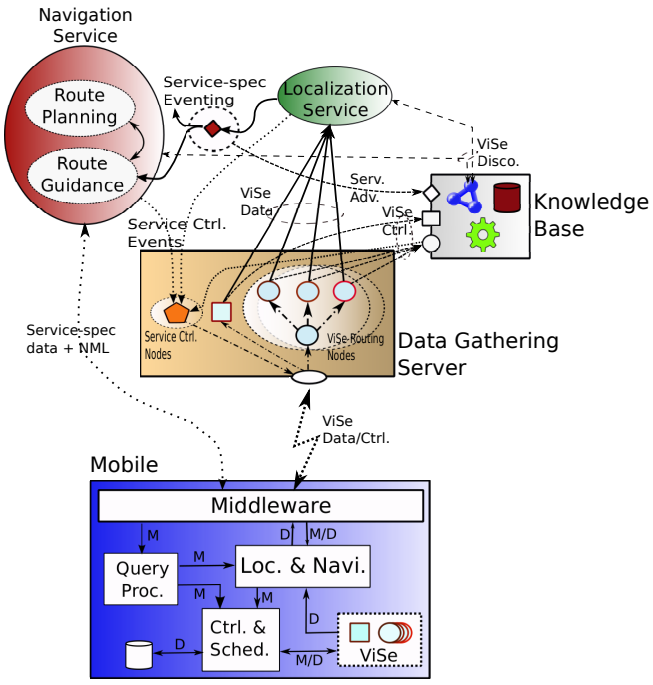


Fig. 2. System architecture of the navigation system

route and triggering the correction procedure is related to the definition of uncertainty areas between two consecutive TAPs (off-line calculation and configuration/setup). This task can be also formulated as an NML query and its result is triggered by an estimation procedure: when the distance between the last know precise location and the estimated actual position exceeds the threshold defined by d_{pen} + d_{rem} (see Figure 1). The refinement in the localization process and the possible recovery of the navigation are calculated based on the reported ViSe data on the server side.

IV. Position Refinement using Cameras

Cameras can be also used as absolute position estimators. Researchers are trying to develop mobile resource effective methods using artificial landmarks (e.g., painted signs [9]) or statistical scene modeling. Unfortunately, these methods are sensitive to the (financial, aesthetic and judiciary) cost of landmarks and the dynamics of the modeled scene. Accordingly, in our case both pre-installed QR codes and 3D reconstruction based positioning will be used as checkpoints. We extract image key-points and calculate 3D coordinates using at least two digital images of the same checkpoint and its surroundings. We use two types of input: normal and panoramic images [12]. If two normal images are used as input, pose (position and orientation) of the camera needs to be recorded. Hence, this method is used by system installers. In case of panoramic images, the pose of the camera can be estimated relative to the reconstruction coordinate system. To correctly align the reconstructed model with the predefined spatial one, arrangement of branches, stairs or exits can be used. Thus, reduction of 3D point cloud can be used by slicing

 ISBN: 978-1-61208-229-5

in the range of typical camera heights. From input images (normal and panoramic) distinctive invariant image features are extracted. Feature vectors are matched across images to triangulate key-point positions. Reconstruction tasks can be done on server or mobile side. During positioning, along the route, the feature vectors and 3D coordinates of a checkpoint are downloaded to the mobile node. Restricted search region of feature extraction and matching is used with respect to the possible camera pose derived from previous pose and inertial sensor data, to reduce computing costs on the mobile device. In order to provide fast localization, we decided to use the (so far) unexploited parallel processing capabilities of mobile GPUs. We are working on models and methods to estimate optimal QR code placements and reconstruction camera poses according to the spatial constraints of the scene, computational and geometric constraints of methods, and inaccuracy models of sensors.

V. Conclusion and Future Work

In this paper, we proposed a novel approach for architectural design of indoor localization and navigation services. Our goal is achieved by employing human behavior analysis to recognize certain movement patterns, such as walking, taking stairs, using elevators, etc. These events are used to define topology specific points, which recognized during the navigation process can suppress the localization error. For complex scenarios with topological black holes (where the distance between two consecutive TAPs is too large) we introduced the concept of SAPs, whose optimal placement can be calculated off-line and their finding can be triggered automatically by the navigation language (NML). The introduction of this special navigation language will let us to control the localization and position refinement process; thus, no RF infrastructure is necessary and continuous user supervision is minimized or at least is automatically triggered by the scenario itself.

As future work, we have to put the pieces together and to implement the localization service on the server side. Currently the data gathering server, the ViSe querying on the mobile side and an off-line version of the movement pattern detection algorithms are available. The camera based position refinement is also under implementation and evaluation.

Acknowledgment

The work of Tamás Helfenbein was founded by the National Innovation Office (NIH) via the project BelAmI_H.

Attila Török was supported from project WayFiS. The project WayFiS no AAL-2010-3-014 has received funding from AAL JP, co-funded by the European Commission and National Funding Authorities of country Spain (MINETUR), Switzerland (OPET) and Hungary (NIH).

References

[1] K. Rehrl, S. Bruntsch and H-J. Mentz, "Assisting Multimodal Travelers: Design and Prototypical Implementation of a Personal Travel Companion," IEEE Transactions on Intelligent Transportation Systems, vol. 8, no. 1, pp. 31-42, March 2007.

[2] Y. Gu, A. Lo and I. Niemegeers, "A Survey of Indoor Positioning Systems for Wireless Personal Networks," IEEE Communications Surveys & Tutorials, vol. 11, no. 1, pp. 13-32, first quarter 2009.

[3] K. Chintalapudi, A.P. Iyer and V.N. Padmanabhan, "Indoor Localization Without the Pain," MobiCom'10, September 20-24, 2010, Chicago, Illinois, USA.

[4] Y. Jin, M. Motani, W-S. Soh and J. Zhang, "SparseTrack: Enhancing Indoor Pedestrian Tracking with Sparse Infrastructure Support," IEEE Infocom 2010, pp. 668-676, NJ, USA, 2010.

[5] Y. Jin, H-S. Toh, W-S. Soh and W-C. Wong, "A Robust Dead-Reckoning Pedestrian Tracking System with Low Cost Sensors," IEEE PerCom, pp. 222 - 230, Seattle, March 21-25, 2011.

[6] Z. Song and et al., "Dead-Reckoning Assisted WiFi Based Indoor Pedestrian Localization," ComNet-IoT, January 3, China, 2012.

[7] M. Löchtefeld, S. Gehring, J. Schning and A. Krüger, "PINwI - Pedestrian Indoor Navigation without Infrastructure," NordiCHI '10: Extending Boundaries, pp. 731-734, New York, USA, 2010.

[8] A. Mulloni, H. Seichter and D. Schmalstieg, "Handheld Augmented Reality Indoor Navigation with Activity-Based Instructions," MobileHCI 2011, pp. 211-220, Aug 30Sept 2, Stockholm, Sweden, 2011.

[9] A. Mulloni, D. Wagner, I. Barakonyi and D. Schmalstieg, "Indoor Positioning and Navigation with Camera Phones," IEEE Pervasive Computing, vol. 8, issue 2, pp. 22-31, 2009.

[10] D. Merico and R. Bisiani, "Indoor Navigation with Minimal Infrastructure," Proc. 4th Workshop Positioning, Navigation and Communication (WPNC 07), pp. 141144. 22-22 March, 2007.

[11] C.W. Han, S.J. Kang and N.S. Kim, "Implementation of HMM-Based Human Activity Recognition Using Single Triaxial Accelerometer," IEICE Trans. Fundamentals, vol. E93A, no.7, pp. 1379-1383, July 2010.

[12] Q. Pan, C. Arth, E. Rosten, G. Reitmayr and T. Drummond, "Rapid Scene Reconstruction on Mobile Phones from Panoramic Images," IEEE ISMAR '11, pp. 55-64, Washington, DC, USA, 2011.

 ISBN: 978-1-61208-229-5

A Smarter Collaborative Mobile Learning Solution

Rui Neves Madeira
DSI, ESTSetúbal,
Politechnic Institute of Setúbal,
Campus do IPS, Setúbal, Portugal
rui.madeira@estsetubal.ips.pt

***Abstract*— Mobile learning has been receiving increased attention from diverse publications and events. The assimilation of mobile computing by education allows students to access education calmly, flexibly and seamlessly. This area of mobile services is getting wide attention by developers, strongly supported by the notion of context-aware. It permits that systems can take the location and position of a user, her interactions and even 'smart' objects into account to present more personalized services. This work presents the addition of a personalization module to PortableLab, a mobile learning solution that allows students to analyze several poor quality power supply occurrences. The developed system is a step forward in the development of mobile learning courses, presenting content adapted to each student. The paper presents the PortableLab system and how its integration of personalization is being done in order to have the right mobile interfaces tailored to the students.**

Keywords-personalization; smart interfaces; adaptation; android; m-learning

I. INTRODUCTION

Nowadays, mobile devices are essential tools for peoples' daily living. The assimilation of Ubiquitous Computing (Ubicomp) [1], strongly based on Mobile Computing, by education, marks an important step forward allowing students to access education calmly, flexibly, and seamlessly. The term Mobile Learning (M-Learning) is frequently used to refer to the use of handheld mobile devices that enable the learner to be 'on the move', providing anytime anywhere access for learning [2]. Moreover, lifelong learning is also a requirement of our era and mobile technologies can help meeting this challenge through the offer of access to "just-in-time knowledge".

M-learning is considered more innovative and student-centered than typical e-learning or classic distance education methods, representing an effective pedagogical method as any other conventional learning method [3, 4]. Furthermore, it is desired that a mobile learning system proactively reacts to individuals who use it, in a pervasive and persistent way. This human-centered vision demands for adaptive and personalized services also according to the context. Personalization must be a major component in m-learning systems, and generally in Ubicomp scenarios. M-learning combined, as much as possible, with other Ubicomp's features can offer great innovation to the learning process, allowing an adaptive learning through personalization according to students' preferences and learning capabilities.

However, this personalization requirement needs more than wirelessly networked computers, sensors and mobile devices working together, as it relies implicitly on some kind of recommendation mechanism to directly serve the individual or the group. The offer of the right personalized content, interfaces and services is a challenge due to various issues such as diverse user interests and particular needs, heterogeneous environments and devices, dynamic user behavior and user privacy. The process of obtaining and choosing relevant content and interfaces for user interaction in m-learning systems is still a critical challenge, being a hard task in many applications from diverse domains [5].

This paper presents PortableLab [6], a mobile laboratory with several interfaces for power quality assessment, giving special attention to the addition of a personalization module, which is still in progress. The main goal of PortableLab is to improve students' interest and motivation, making resources available as much as possible at any place and any time. This mobile remote laboratory is being used as a complement to the usual classroom laboratory type lessons. The developed system integrates a server with a data acquisition board and a central database to be accessed by the mobile application, programmed for the Google Android platform. The mobile application includes a collaborative learning module that it is essential for the growth of students. With this module, they can annotate content to be seen by teachers and colleagues, giving additional information about their understandings or helping others in the learning process. Furthermore, we are in the process of integrating a module responsible for adapting contents and functionalities to the level demonstrated by students and also according to the interactions stream, which is defined by screens and components clicks and functionalities executed by users.

The paper is organized as follows. In Section II, we present a summary of related work. The third Section introduces the system's architecture and the mobile application interfaces. Section IV adds the ideas about the personalization process and smart adaptation of PortableLab interfaces. Finally, in Section V, conclusions and future work are presented.

II. RELATED WORK

Nowadays, m-learning is a very active research field, with the development of many important and interesting projects. M-learning is rapidly growing from a set of research projects into worldwide deployment of services for

 ISBN: 978-1-61208-229-5

classrooms, field trips, workplace training and informal education, among other areas.

Frohberg et al. presented a deep and critical analysis of m-learning projects published before the end of 2007 [7], yet without important focus on personalization to tailor content and interfaces to the right student and even teacher. Major m-learning projects have been concentrating on the generic platforms development for m-learning and explored new supports for a kind of technology-mediated learning across locations and life transitions [8, 9].

Smaller projects are more directed to develop new pedagogical solutions for specific cases and to explore how learning on handheld mobile devices interweaves with personal interests and individual learning needs [10], which is much more of our interest. The SHAPE project [11] is one that can benefit from this approach in a next phase. It aims to enhancing the conceptual understanding of how to undertake design of computing in public spaces and to create exemplars for how new computing can be used to augment educational and social interaction in public environments. A part of the project is used to simulate an archaeology dig, where the aim is to enhance children's collaborative learning in museums, through supporting sensorial experience and capturing embodied knowledge. It can be tailored to each user. UNIWAP [12] is another interesting m-learning project created to assist in teacher training. The project used relatively simple technologies, short message service (SMS) and digital pictures, to enable students to create digital portfolios built from materials created in the field. Messaging was used to enable the trainees, who were widely distributed when training in different schools, to collaborate with each other and share their experiences. These projects are interesting because of the collaborative approach they include, presenting some personalization possibilities.

Finally, Kay presents a "vision for the lifelong user model as a first class citizen, existing independently of any single application and controlled by the learner" [13]. She argues that it has a key role for a vision of personalized lifelong learning, enabling students/learners to supplement their own knowledge with readily accessible digital information based on information they have accessed or used. The paper also presents a good overview of Intelligent Tutoring Systems, which are important in terms of features related to adaptation.

Although being possible to find m-learning projects using personalization or adaptation concepts, it is clear that none uses a general model based on web-services platform that integrates machine learning modules like the approach applied to PortableLab.

III. PORTABLELAB

This Section is used to present PortableLab, its architecture and technologies, also introducing the main screens in order to understand where personalization can be applied.

A. *Implementation of PortableLab*

The architecture of PortableLab can be seen as a typical client-server approach. The data acquisition is done by current and voltage sensors that send signals to a data acquisition board, which is connected to a server that runs signals processing and data management modules. These modules store the received data in the server's main database. This database is updated every time new values are read, independently of the requests of the mobile clients. Apart from the existence of the server database (remote to the mobile user), the user can also choose to use a (local) database, located in the mobile learning application. To access the remote database, mobile devices need to use a PHP (PHP: Hypertext Preprocessor) API (Application Programming Interface) through HTTP (Hypertext Transfer Protocol) connections to obtain the necessary information for the reproduction of various types of charts, including harmonic content, voltage, and current charts. This API is also used to synchronize the information between the two databases, with the communication being made in a bidirectional way. The master database is always the remote database, since it will store the real-time measured data. Its content is copied to the mobile database every time the user chooses to synchronize both databases. The main programming language of the mobile application interfaces is Java, mostly using libraries provided by the Android SDK and the AChartEngine library that provides graphical functionalities to reproduce the various types of charts. A detailed description of the system and its technologies can be found in [6].

B. *PortableLab's Interfaces*

The Login screen where the user, a student or teacher, has to authenticate is the entrance of the application (Fig. 1, left image). After a successful login, the user finds the main screen where s/he can choose the visualization of charts with the most recent reads related with power quality (Fig. 1, right image at bottom). Another functionality included in this screen is the search of reads by date, time, or by both.

Figure 1. Mobile application's initial screens (in Portuguese).

The left image in Fig. 2 presents how the PortableLab application appears in a real device in terms of representation of charts, and the right image shows below the charts the comments/annotations made by students and teachers,

illustrating the collaborative role of the tool. A student can use this functionality to annotate some results, put questions to teachers, or help other students. A teacher can use it to respond to students or to better explain an idea about the results.

Figure 2. Mobile application's screens on a HTC device: charts visualization and list of recent annotations by users (in Portuguese).

The application presents other functionalities, such as, the application's operation mode configurations, the user personal data and the user authentication definitions. Depending on the user profile, other options are available, such as the possibility of seeing a detailed list of users that have used the system. Only teachers have access to this screen.

IV. INTEGRATION OF PERSONALIZATION

This Section discusses how to apply personalization, smart adaptation, to Ubicomp and Mobile Computing systems, presenting a general solution. Moreover, it presents initial decisions regarding PortableLab's personalization integration.

A. A Web-based Personalization Model Solution

An essential input for every intelligent adaptation or personalization technique is the user model [13]. However, it is important to define the information integration level since the focus should be on the integration of personal (from the user model) and contextual information (from the context) about the user in the current domain of application. Additionally, the adoption of ontology for smoothly modeling the domain, the context and the personalization process can contribute to tailor the right information, services or interfaces to users, thus, facilitating and enriching the HCI (Human-Computer Interaction) process. Ontologies can also be very important for the reuse of parts of the user model in a ubiquitous environment. This requires protocols for ontology understood by applications and/or a mechanism for mapping different ontologies within them [13]. At the moment, we are working on the ontology definition as we consider it very important to obtain the best personalization model. We are also developing the user and context models that are part of the latter.

A partial taxonomy of what we consider the starting point for user modeling integrates: demographics, preferences, roles, and knowledge. In terms of intelligent adaptation of the interfaces and content, it will work much better if besides the usual user profile we consider the location, both physical and semantic (e.g., at home), the situation (e.g., alone, in family or lunching) and the emotions felt by the user (initially, s/he can choose from a list), connecting to the context model. Context awareness plays a major role in Ubicomp, being tightly coupled with user modeling [14]. Moreover, an interactions stream (set of interactions between the user and the application) is used to know the "degree of empathy" between the user and the system/application.

So, the personalization model is based on a generic configuration data model that is mainly composed of Personalization Options, Parameters and Resources (see Fig. 3). Data concerning direct user interaction with the application, such as clicks, time spent on menus, and numbers of log-in operations are considered as Resource data. Parameters are usually defined by mathematical expressions based on the Resources (seen as variables) and used to characterize the different options for each desired personalization (see Table 1 for an example). This configuration model is also a part of the general personalization model, being closely connected to the user and context sub-models.

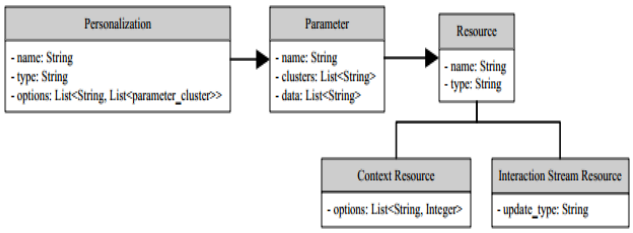


Figure 3. The personalization configuration model.

When interacting with an application, users do not behave in a uniform way. Different users may present different interests, different knowledge and different capabilities to interact with the system. Therefore, it is important to assign users to categories that represent them to facilitate the personalization process. It is possible to use the interaction stream data, in conjunction with user and context (if considered) data, in a clustering operation with a defined number of clusters, each one basically representing a user category. The result of this operation is a set of categories (clusters) where each one contains a collection of users (points). After the clustering operation, a recommendation service checks which personalization options are associated to the user categories in order to link the options to the users. Moreover, we are still working on the use of Recommender systems (RS) to improve personalization, adding levels of detail to the machine learning (ML) algorithm.

 ISBN: 978-1-61208-229-5

RS can be seen as mediators of the user experience in the digital world and are increasingly helpful in doing the same in the physical world [15]. The goal of a common RS is to proactively suggest and prioritize items the user may be interested in, taking into account the context at the moment of the interaction, and predicting the user behavior [16]. The application of context to RS can be tightly integrated into the recommendation algorithm, or used independently to improve its recommendations. Based on the general personalization model, integrating user, context, application and domain, and configurations sub-models, candidate items such as interface customization and learning information are selected to feed up the ML algorithm, which we propose with a composition of four main sub-modules (Fig. 4).

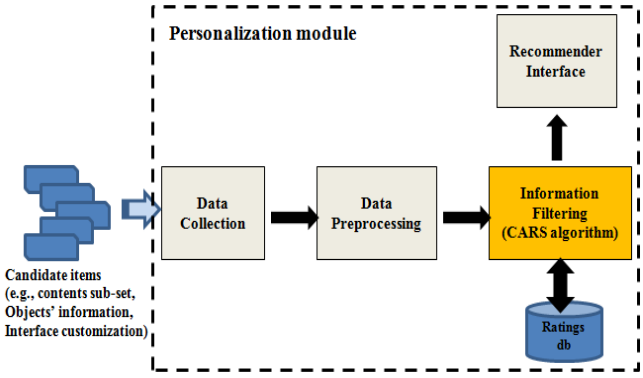


Figure 4. Architecture of modules involving the recommendation process.

Being based on RS, the personalization module needs a ratings database, also based on feedback provided by users and on their interactions stream. In an m-learning system, based on Ubicomp characteristics, the goal is to have a system that is as unobtrusive as possible. So, how can the system gather the ratings from users, requiring as little interaction as possible? The rating can be explicit when in specific moments the system asks for a simple rating of an interface or some piece of information. More interesting are the implicit ratings, which can be inferred from the interactions of the user with the system. It can be based on the time spent in front of the display, analyzing face expressions, or more simply on number of screens and components clicks and functionalities executed by users.

Furthermore, we intend to have this general personalization model being applied to different applications, systems, even if from different domains. The model is deployed as a framework, a web-services platform – personaX - designed to provide orientation and tools to help developers in the implementation of a standard personalization. On the other side, users will be less bothered when starting to use a new system/app. This one already might know something about s/he. User interactions within one system might be useful to help personalizing another one.

The core of personaX (Fig. 5) is working, being used in the form of personalization APIs and configuration modules, which give the developer a high-level of implementation. The personalization algorithms are already implemented and the developer only needs to apply the model according to specific project's needs.

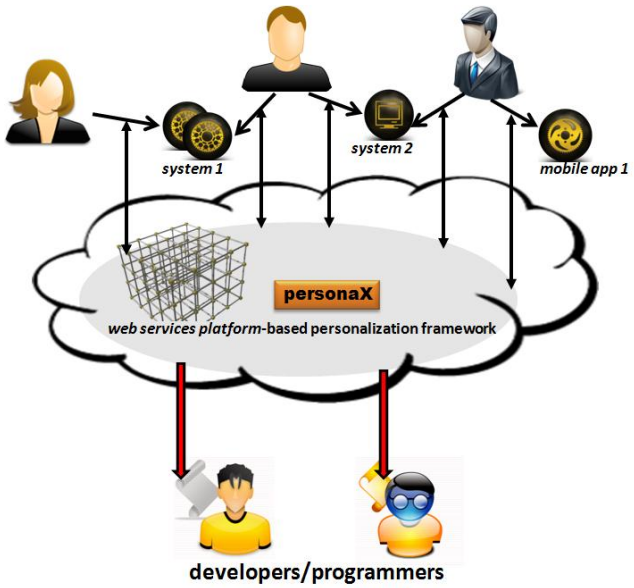


Figure 5. personaX: a web-services platform for personalizatin.

B. *PortableLab's Initial Personalization Decisions*

Furthermore, we have to decide about the user data that can and should be captured and included as a basis for the personalization. To accomplish that capture, which are the potential sources for user modeling information?

The data can be collected from, for instance: life logging sensors, personal devices and in the cloud, besides the specific Information System of the University using PortableLab. With more sensors, such as GPS, camera, and gyroscope, embedded in mobile devices, it is possible to record the activities and behavior of a user. Additionally, the usage of wearable sensors enables the capture of users' physical and physiological data and, this way, the emotional model [15] can be predicted and not only stated by the user. The captured sensor data can be used to represent a situation event and extract patterns from the activities data logs. In relation to context, the environment conditions can be easily inferred with the usage of sensors, while the location information can be known using GPS and Wi-Fi technologies.

TABLE I. EXAMPLE OF PERSONALIZATION OPTIONS FOR PORTABLELAB

Personalization	***Personalization Options***	**Parameters**	**Resources**
Initial screen	- Communicator - Solitary	Communication profile: - numberPosts /numberVisual	-numberPosts -numberVisual

V. CONCLUSIONS AND FUTURE WORK

The use of m-learning tools, if correctly contextualized and built, can benefit the traditional learning methodologies and methods. However, these m-learning applications should integrate a personalization approach to have better chances of being really efficient working as complementary learning tools. This work focuses on a proposal for facilitating the implementation of personalized m-learning systems.

PortableLab is being tested as a first m-learning prototype with an explicit personalization module. The research is studying how the personalization model can be defined, distributed and executed among different devices. These personalization features are currently being defined in order to choose the best ML algorithms, along with the definition of the final user and context models. The applicability and maturity of the proposal should be shown through its usage in the development of the prototype.

ACKNOWLEDGMENT

I would like to thank my advisor Prof. Dr. Nuno Correia for supporting me on the personalization studies, Marco Guerra and Cláudia Francisco for previous work on the PortableLab development and, finally, André Vieira for debating ideas about the personalization model.

REFERENCES

[1] M. Weiser, The Computer for the 21st Century, Scientific American, vol. 265, no. 3, Sept. 1991, pp. 66-75.

[2] S. Price, Ubiquitous computing: digital augmentation and learning, in Pachler, N. (Ed.), M-learning: towards a research agenda, WLE Centre occasional papers in work-based learning 1, London: Institute of Education, 2007.

[3] C. Romero, S. Ventura, and P. Bra, Using Mobile and Web- based Computerized Tests to Evaluate University Students, Computer Applications in Engineering Education Journal, vol. 9999, Published online in Wiley InterScience, 2009.

[4] R.N. Madeira, V.F. Pires, O.P. Dias, and J.F. Martins, Development of a Mobile Learning Framework for an Analog Electronics Course, Proc. Int. Conf. IEEE Education Engineering (EDUCON), Madrid, Spain, 2010, pp. 561 – 567.

[5] M. Gorgoglione, C. Palmisano, and A. Tuzhilin, Personalization in context: Does context matter when building personalized customer models?. In ICDM '06 , pp. 222–231, 2006.

[6] M.A. Guerra, C.M. Francisco, and R.N. Madeira, PortableLab: Implementation of a mobile remote laboratory for the Android platform, In Global Engineering Education Conference (EDUCON), 2011 IEEE, pp.983-989, April 2011.

[7] D. Frohberg, C. Göth, and G. Schwabe, Mobile Learning projects – a critical analysis of the state of the art. Journal of Computer Assisted Learning, vol. 25, no. 2009, pp. 307–331.

[8] C. H. Muntean, and G. Muntean, Open Corpus Architecture for Personalised Ubiquitous E-learning. Personal Ubiquitous Computing journal, vol. 13, no. 3, 2009, pp. 197-205.

[9] M. Sharples, The Design of Personal Mobile Technologies for Lifelong Learning, Computers and Education, vol. 34, no. 3,2000, pp. 177-193.

[10] A. Hamid, and S. Hafizah, WE-learning – Electronic and Mobile Learning Environment, The Public Institutions of Higher Learning R&D Exposition, Kuala Lumpur, Malaysia, 2003.

[11] T. Hall, and L. Bannon, Designing ubiquitous computing to enhance children's interactions in museums. Proceedings of IDC 2005, Boulder, Colorado, 2005.

[12] P. Seppälä, and H. Alamäki, Mobile learning in teacher training, Journal of Computer Assisted Learning, No. 19, 2003, pp. 330-335.

[13] J. Kay, Lifelong Learner Modeling for Lifelong Personalized Pervasive Learning. IEEE Trans. Learn. Technol. 1, 4, 2008, pp. 215-228.

[14] A. Jameson, and A. Kruger, Preface to the Special Issue on User Modeling in Ubiquitous Computing. User Modeling and User-Adapted Interaction 15, 3-4 (August 2005), pp. 193-195.

[15] McDonald D.W.: Ubiquitous Recommendation Systems. IEEE Computer, pp 111-112, 2003.

[16] Palmisano C., Tuzhilin A., Gorgoglione M.: Using context to improve predictive modeling of customers in personalization applications. In IEEE Trans. On Knowl. and Data Engineering, 20(11), pp. 535–1549, 2008.

[17] González G., López B., Rosa J.L.: Managing Emotions in Smart User Models for Recommender Systems. In Proc ICEIS'04, pp. 187-194, 2004.

 ISBN: 978-1-61208-229-5

Optimized Flow Management using Linear Programming in Future Wireless Networks

Umar Toseef*†, Andreas Timm-Giel*
*Institute of Communication Networks
Hamburg University of Technology
Schwarzenbergstr. 95E, 21073 Hamburg, Germany
Email: {umar.toseef, timm-giel}@tuhh.de

Carmelita Görg†
†TZI ComNets - University of Bremen
Otto-Hahn-Allee, NW1
Bremen, Germany
Email: {umar,cg}@tzi.de

***Abstract*—There have been tremendous advances over the past decades when it comes to wireless access technologies. Nowadays, mobile devices are equipped with several wireless access technologies like 3G, 4G or WiFi. Currently, these mobile devices can communicate using one access technology at a time. However, there is a big potential for improving network capacity and enhancing user 'Quality of Experience' if these access technologies are integrated. Such an integration would allow access technologies to cooperate and work simultaneously in a heterogeneous environment from which both the end users as well as the mobile operators can benefit. In this paper, it is investigated how to tackle the simultaneous usage of wireless access technologies. For this purpose, a practical example of a 3GPP LTE and a non-3GPP WLAN integrated heterogeneous network is considered. Furthermore, a novel decision mechanism is proposed, that focuses on optimizing the flow management of user traffic flows based on a mathematical formulation of the system. The mathematical model is implemented using Linear Programming techniques. The paper demonstrates the gains and benefits that are achieved from using such innovative decision mechanism as well as the benefits that arise from the simultaneous usage of wireless heterogeneous accesses.**

Keywords: ***LTE and WLAN, Resource Allocation, User QoE, Heterogeneous Networks, Linear Programming.***

I. INTRODUCTION

There are various prevailing standards of wireless access technologies in the current communication market, such as 3GPP (3rd Generation Partnership Project), non-3GPP, 3GPP2, etc. Admittedly, each type of these access technologies has certain advantages, which justify its existence in this age of evolution of technology. For example, 3GPP networks are more efficient in terms of handling high traffic demands, providing QoS (Quality of Service) guarantees and the extended coverage. Whereas, the non-3GPP technologies like IEEE 802.11 [2] are simple to operate and therefore need less investment and operation & maintenance cost. On the other hand, the wireless portable devices are becoming increasingly popular and it is widely expected that such devices will outnumber any other forms of smart computing and communication in near future. With the capability of connecting through several types of 3GPP and non-3GPP access technologies these devices run a wide variety of bandwidth demanding services including high speed data delivery and multimedia communication. However, due to the limitations of today's network architecture these devices can connect to one access technology a time. It is proposed that in future networks of heterogeneous access technologies, the ever increasing bandwidth demands of the portable devices can be better addressed through the bandwidth resource aggregation of multiple networks. This would create a win-win situation for the network operators and the users. 3GPP standardization has already envisioned the possible benefits from the cooperation of 3GPP and non-3GPP networks and has come up with such integration standards[5]. 3GPP specified System Architecture Evolution (SAE) allows mobile users to roam between 3GPP and non-3GPP access technologies with seamless mobility provided through Proxy Mobile IPv6 (network based mobility) and Dual Stack Mobile IPv6 (host based mobility) [5]. The 3GPP SAE architecture, however, does not support the user multi-homing i.e. simultaneous user connection to more than one access network. In order to investigate the achievable advantages through the support of user multi-homing, the existing 3GPP standards needs extensions. Moreover, the issues related to an efficient management of aggregated bandwidth resource should also be addressed when multi-homing support is realized.

The focus of this work is to extend the 3GPP SAE architecture in realizing multi-homing support for users and propose a solution to make an optimum use of aggregated bandwidth resources and network diversity in a multi-homing scenario.

The rest of the paper is organized as follows: Section II describes how the current 3GPP SAE architecture is extended to provide users with multi-homing support. Section III describes the importance of flow management function in a heterogeneous network, and Section IV explains the linear programming technique used to achieve an optimized flow management operation. Finally, Section V provides the proof of concepts through the discussion of simulation results of an investigated realistic scenario.

II. NETWORK SIMULATION MODEL

This work follows the proposal of 3GPP specifications in the integration of 3GPP access technology (namely LTE) and trusted non-3GPP access technology (namely legacy WLAN 802.11g), where host based mobility solutions, i.e., Dual Stack Mobile IPv6 is considered. For this purpose, a simulation network model has been implemented using the OPNET [6]

 ISBN: 978-1-61208-229-5

network simulator. This includes the detailed implementation of LTE network entities following the 3GPP specifications [8]. Simulation models of WLAN access points as well as, the common protocol layers like application, TCP/UDP, IP, Mobile IP, Ethernet, etc., come from the OPNET standard library [6]. The home agent (HA) function is located at the Packet Data Network (PDN) gateway. The remote server acts as a correspondent node (CN) from where mobile users access application services; (see Fig. 4). A user can have up to two active network interfaces, one for each access technology. Further details about the simulator can be found in [2].

III. Flow Management

Flow management helps a network operator or a user to make use of the two network paths from available access technologies. In general, there are two options of managing traffic flows for a multi-homed user. The first option is to carry one complete application traffic flow over one path of choice, this is known as "traffic flow switching". For example, a user can decide to keep his TCP based traffic flows on the WLAN access network while his VoIP/video traffic follows its way over the LTE network. The second option is to divide the traffic flow into several smaller sub-flows where each sub-flow is carried over one network path. This will be called "traffic flow splitting". For example, a user watching HD video streaming of a football match can distribute the video traffic flow over the WLAN and LTE network as long as the user is in the overlapped coverage of both networks. In this work, both of these options will be used based on the requirements of global optimal resource allocation goal.

In a wireless access network, frequency spectrum and its usage time are the main network resources, which are shared by all users. If a user has good channel conditions, he can use higher modulation schemes and achieve higher spectral efficiency. High spectral efficiency allows the user to transmit more data bits for a given amount of network resources. The opposite is true for a user who is suffering from bad channel conditions. The amount of network resources of a wireless network are determined from the designed parameters like frequency spectrum, transmission technology, antenna gains etc. Therefore, each network has a fixed amount of network resources and the network performance itself depends on the fact with what spectral efficiency these resources are utilized. In order to achieve higher data rates a network resource scheduler should select those users who can attain high spectral efficiency, and therefore, need less network resources per unit data rate. In this work, we adapt the term "network path cost" for the required network resources per unit data rate. For a user, its network path cost can be accessed through cross layer information from the MAC layer of the corresponding access technologies. In the following subsections, it is shown how the network path cost can be computed for users in LTE network and how the achievable user throughput can be estimated in WLAN network.

A. Network path cost for LTE

LTE performs a managed scheduling of available bandwidth resources. The smallest unit of bandwidth resource is referred to as a physical resource block (PRB) in the LTE specification. Based on the allocated frequency spectrum size LTE has a certain number of PRBs. The LTE MAC scheduler residing at the eNodeB schedules these PRBs using a 1ms transmission time interval (TTI). The LTE MAC scheduler has a very complex way to assign resources to the associated users. Without digging into the details of the MAC scheduler operation, we focus on the last stage of resource assignment procedure in a certain TTI. On reaching that stage the MAC scheduler already builds up a list of users, which will be transmitting/receiving data in that TTI. For each user entry in the list, there is a corresponding value of the allocated number of PRBs, as well as the channel dependent Modulation and Coding Scheme (MCS) index. These two values are used to lookup the Transport Block Size (TBS) from a table defined in the 3GPP specifications [1]. This is a two dimensional table where each row representing one MCS index lists several values of TBS corresponding to the allocated number of PRBs. The obtained TBS value defines the size of the MAC frame transmitted to the user in that TTI. In this way, the user received throughput at the MAC layer in a certain TTI can be estimated if the TBS value for that user is known.

Fig. 1(a) shows that for a particular MCS index, the LTE throughput value has almost a linear relationship with the used number of PRBs. If described mathematically this relationship can be used to determine the required number of PRBs/TTI (q) to achieve a certain data rate R_i [kbit/sec] for a user having MCS index i. That is

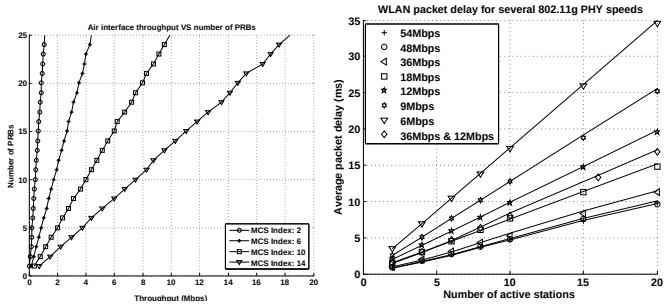
$$q = \alpha_i \cdot R_i + \beta_i$$

α_i is slope of a straight line (as shown in 1(a)) described in units of PRBs/kbps. β_i is the intercept at the y-axis and has units of number of PRBs. It can be noticed that α_i is the data rate dependent part, while β_i is the data rate independent part of network resource requirement for a user with channel conditions mapped to MCS index i.

B. Average user throughput estimation in WLAN network

The user throughput in WLAN (IEEE 802.11) can be computed if the packet transmission delay is known. However, due to the random back-off time and possible packet collisions, the time required to transmit a packet successfully is highly variable. Moreover, this transmission time also depends on several factors like user channel conditions, number of users, user traffic demands etc. In [10], two dimensional Markov chain model has been used to compute the achievable throughput of 802.11b network with a certain number of stations having same channel conditions and traffic pattern. [11] has extended the model to calculate the average packet delay. The mathematical analysis in [11] assumes a network of n contending stations where each station has always a packet to transmit. The analysis yields two probability value: probability that there is at least one transmission in the considered

 ISBN: 978-1-61208-229-5



(a) (b)

Fig. 1. The left figure shows the relationship of LTE air interface user throughput and number of PRBs for different MSC index values. Each curve represents one MCS index. The right figure shows the packet delay estimation for users in WLAN network using proposed analytical model.

slot time (P_{tr}) and the probability that an occurring packet transmission is successful (P_s).

$$P_{tr} = 1 - (1-\tau)^n, \qquad P_s = \frac{n \cdot \tau \cdot (1-\tau)^{n-1}}{1-(1-\tau)^n}$$

where τ is the stationary probability that the station transmits a packet in a randomly chosen slot time. The relationships of P_{tr} and P_s are used to calculate the $E[slot]$ which is the average length of a slot time. The average length of a slot time is obtained considering that, with probability $1 - P_{tr}$, the slot time is empty; with probability $P_{tr}P_s$ it contains a successful transmission, and with probability $P_{tr}(1 - P_s)$ it contains a collision.

$$E[slot] = (1-P_{tr})\sigma + P_{tr} \cdot P_s \cdot T_s + P_{tr}(1-P_s) \cdot T_c \quad (1)$$

here σ is the duration of an empty slot time, T_s is the average time the channel is sensed busy because of a successful transmission, and T_c is the average time the channel is sensed busy by each station during a collision. Assuming $E[X]$ as the average number of slot times for a successful packet transmission, the value $E[X]$ can be found by multiplying the number of slot times the packet is delayed in each back-off stage by the probability to reach this back-off stage. The final form of $E[X]$ is as given below

$$E[X] = \frac{(1-2p) \cdot (W+1) + pW \cdot (1-(2p)^m)}{2 \cdot (1-2p) \cdot (1-p)} \quad (2)$$

Finally, the average delay of a successfully transmitted packet $E[D]$ is given as following

$$E[D] = E[X] \cdot E[slot] \quad (3)$$

In equation (3), the values of σ, m and W can be obtained from the 802.11 specifications. The values of the other two unknown parameters i.e., T_s and T_c depend on fact whether the basic or RTS/CTS scheme is used. For example, with RTS/CTS scheme enabled these values are as follows

$$T_s^{rts} = T_{RTS} + T_{SIFS} + \delta + T_{CTS} + T_{SIFS} + \delta + T_H + T_{E[P]} + T_{SIFS} + \delta + T_{ACK} + T_{DIFS} + \delta \quad (4a)$$

$$T_c^{rts} = T_{RTS} + T_{DIFS} + \delta \quad (4b)$$

where δ is the propagation delay and $T_H = T_{PHYhdr} + T_{MAChdr}$ is the time to transmit header data associated with PHY and MAC protocols and $T_{E[P]}$ is the time to transmit a data packet of mean size $E[P]$.

In the above described analysis, it has been assumed that all stations have same the channel conditions and therefore transmit with the same PHY data rate. In a realistic scenario this assumption cannot always be fulfilled. In order to use equation (3) in a scenario where users have different channel conditions and PHY data rates, it must be extended. The direct influence of the PHY data rate on average packet delay estimation can be observed in the computation of T_s and T_c (see equation (4)) where the user PHY data rate determines the value of $T_{E[P]}$, the time to transmit the data packet. Therefore, a network of users with different PHY data rate can be seen as a system with single server and multiple queues where the user PHY data rate is incorporated in the size of job. Excluding the medium contention time and assuming that a user always has a packet to transmit, the mean service time of such a system can be computed as

$$\hat{T}_{E[P]} = \frac{E[P]}{E[\text{PHY data rate}]}$$

Furthermore, when stations are transmitting at different PHY data rate, the transmission speed of control signals in the network, i.e., T_H, T_{RTS}, T_{CTS} and T_{ACK} is limited by the station having the lowest PHY data rate. This implies that the users must transmit control signals at a PHY data rate, which can be received by the all users. But, the data packet is transmitted by the user's own current PHY data rate. Incorporating the modified values of T_s and T_c in equation (1) produces $\widehat{E[slot]}$

$$\widehat{E[slot]} = (1-P_{tr})\sigma + P_{tr}P_s\hat{T}_s + P_{tr}(1-P_s)\hat{T}_c \quad (5)$$

As the value of $E[X]$ is independent of user PHY data rate, equation (3) takes the following form

$$\widehat{E[D]} = \widehat{E[slot]} \cdot E[X] \quad (6)$$

Figure 1(b) shows the average packet delay experienced by users transmitting in 802.11g network with RTS/CTS enabled. The solid lines show the estimated values using the equation (6). The markers on a solid line represents the delay values obtained from simulation results. It is evident from the figure that the modified model can precisely estimate the mean packet delay values for both scenarios, i.e., when all users have same PHY data rate as well as when users with different PHY data rate are mixed together. The mean packet delay value computed with (6) can be used to estimate the average user throughput Y in the network, i.e.,

$$Y = \frac{E[P]}{\widehat{E[D]}} = \frac{E[P]}{E[X] \cdot \widehat{E[slot]}} \quad (7)$$

 ISBN: 978-1-61208-229-5

IV. Optimized network resource allocation

When the network path costs for a multi-homed user are known, the problem of optimal resource utilization can be solved using mathematical techniques. In this work we have selected Integer Linear Programming (ILP) to solve this problem. A mathematical model for this purpose is discussed in first subsection. The second section explains how the non-linear relation for WLAN throughput computation is linearized and the third section shows how this model is integrated into simulation environment.

Given

U	a set of users
α_j	Data rate dependent part of the LTE link cost in PRBs per kbps for user j, for each $j \in U$
β_j	Data rate independent part of the LTE link cost in PRBs for user j, for each $j \in U$
ϕ_j	WLAN PHY data rate of a user j, for each $j \in U$
λ_j	Minimum data rate (kbps) demand of a traffic flow destined to user j, for each $j \in U$
Λ_j	Maximum data rate (kbps) allocation for a traffic flow destined to user j, for each $j \in U$
Ω	Number of available PRBs for the LTE access network
G	Mean packet size of active WLAN users in bit
$\widetilde{T_j}$	Per bit transmission delay excluding medium contentions for a user j with PHY data rate ϕ_j, for each $j \in U$

Defined variables

R_j	Size of sub-flow in kbps sent over the LTE access link to user j, for each $j \in U$
V_j	Size of sub-flow in kbps sent over WLAN access link to user j, for each $j \in U$
Y	Average throughput of active users in WLAN network in kbps
E_j	Auxiliary binary variable; its value for a user j is either 1 if $R_j > 0$ or 0 otherwise, for each $j \in U$
F_j	Auxiliary binary variable; its value for a user j is either 1 if $V_j > 0$ or 0 otherwise, for each $j \in U$

Maximize

$\sum_{j \in U} R_j + Y$

Subject to

1. $\sum_{j \in U} (\alpha_j \cdot R_j + \beta_j \cdot E_j) \leq \Omega$
2. $\lambda_j \leq R_j + V_j \leq \Lambda_j, \quad \forall j \in U$
3. $\sum_{j \in U} F_j \geq 1$
4. $\sum_{j \in U} F_j \cdot V_j = \sum_{j \in U} F_j \cdot Y$
5. $V_j \leq G/\widetilde{T_j}, \quad \forall j \in U$
6. $0 \leq R_j \leq \Lambda_j, \quad \forall j \in U$
7. $0 \leq Y_j \leq \Lambda_j, \quad \forall j \in U$

Fig. 2. Mathematical model for the resource allocation in algebraic form

A. Mathematical model for resource allocation

Fig. 2 shows the formulation of the problem in algebraic form. The model defines U as the set of multi-homed users. Each element of this set has a number of input parameters, e.g., network path cost for LTE (α,β) and user WLAN PHY data rate (ϕ) according to the user channel conditions in the corresponding network. The maximum and minimum range of user data rate demands (λ, Λ), which is based on the individual user application. The amount of available network resources in LTE (Ω) are also considered as input parameters. The output parameters for each user in set U include the assigned data rate over the LTE network and the WLAN network paths (R,V). It is obvious that the goal of this model is to achieve the highest possible spectral efficiency from the two network access technologies. The higher the spectral efficiency, the higher the network throughput. Hence the objective is to maximize the user data rate over the two network paths, i.e., R and V for every multi-homed user. The model imposes several constraints, however for the sake brevity the most important seven constraints are listed in Fig.2. The first constraint ensure that the available LTE network resources should not be exceeded when allocating the data rates for users. The second constraint dictates that the user data rate allocation should lie within the specified range. The third constrains enforce the use of WLAN network by at least one user. The 4th constraint allows the users to distribute the available WLAN network throughput according to their needs. This means if a user does not always have some data to transmit his throughput share can be used by other users. Nevertheless, a user cannot transmit at higher data rates than allowed by his PHY data rate as seen in constraint 5. According to the requirements of goal of optimized resource allocation a user may receive multiple sub-flows or one single flow of application data as shown in constraint number 6 and 7.

B. Linearizing WLAN throughput estimation formula

The mathematical model has to rely on equation (7) for the estimation of the average user throughput in WLAN network. However, it is clearly a nonlinear relation, which must be linearized using some work around to use in linear programming. For this purpose, the equation (5) is split into two parts. One part depends only on n, which represents the total number of active stations in WLAN network. The other part incorporates both n as well as $\hat{T}_s$ variables. As RTS/CTS is enabled for all users therefore $\hat{T}_c = T_c$.

$$\widehat{E[slot]} = f_x(n) + f_y(n) \cdot \hat{T}_s$$

where $f_x(n) = (1-P_{tr})\sigma + P_{tr}(1-P_s)\cdot T_c, \quad f_y = P_{tr} \cdot P_s$

Moreover, equation (2) shows that $E[X]$ is a function of only one variable n, which allows us to write

$$\widehat{E[D]} = E[X] \cdot \widehat{E[slot]} = f_1(n) + f_2(n) \cdot \hat{T}_s \qquad (8)$$

where $f_1(n) = E[X] \cdot f_x(n)$, and $f_2(n) = E[X] \cdot f_y(n)$

In order to simplify the relation in equation (8), $f_1(n)$ and $f_2(n)$ are approximated using 3rd order polynomial curve fitting as shown below.

$$f_1(n) \approx A_{11}n^3 + A_{12}n^2 + A_{13}n + A_{14},$$
$$f_2(n) \approx A_{21}n^3 + A_{22}n^2 + A_{23}n + A_{24}$$

where all occurrences of A represent constant value numbers. Fig. 3 shows that the curve fitting process generates an

 ISBN: 978-1-61208-229-5

accurate enough approximation for $f_1(n)$ & $f_2(n)$ functions with norm of residuals as 3.9×10^{-5} and 0.14 respectively. Using the approximate function of $f_1(n)$ and $f_2(n)$ in equation

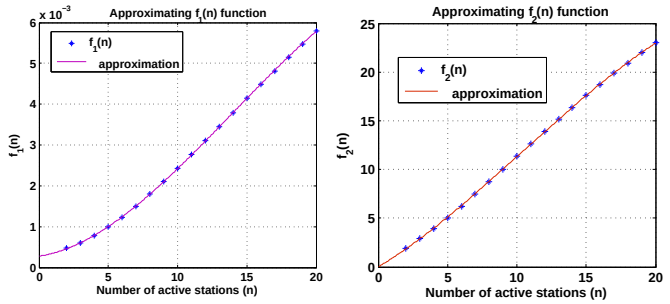


Fig. 3. Approximation of $f_1(n)$ and $f_2(n)$ using polynomial curve fitting

(8) & (7) and after a few manipulation steps of algebra we get

$$n \cdot G = Y \cdot \{A_{11}n^4 + A_{12}n^3 + A_{13}n^2 + A_{14}n + [A_{21}n^3 + A_{22}n^2 + A_{23}n + A_{24}] \cdot \sum_{i=1}^{n} T_{s_i}\} \quad (9)$$

The variable n in above equation can be replaced with a summation of binary variables F_i, which represents whether a station i is active or not. If there are total Z number of users in WLAN network out of which only n users are active then $n = \sum_{i=1}^{Z} F_i$. Similarly higher order variables of n can linearized as following

$$n^2 = (\sum_{i=1}^{Z} F_i)^2 = \sum_{i,j=1}^{Z} F_i \cdot F_j = \sum_{i,j=1}^{Z} \chi_{i,j}^{\text{F2}}$$

Where $\chi_{i,j}^{\text{F2}}$ represents a product of two binary variables F_i & F_j. The value of $\chi_{i,j}^{\text{F2}}$ is determined by following three constraints.

$$\chi_{i,j}^{\text{F2}} \leq F_i, \quad \chi_{i,j}^{\text{F2}} \leq F_j, \quad \chi_{i,j}^{\text{F2}} \geq F_i + F_j - 1$$

Moreover, the continuous variable Y can be multiplied with binary product variable $\chi_{i,j}^{\text{F2}}$ to get the product term $\chi_{i,j}^{\text{YF2}}$, i.e.,

$$Y \cdot n^2 = Y \cdot \sum_{i,j=1}^{Z} \chi_{i,j}^{\text{F2}} = \sum_{i,j=1}^{Z} \chi_{i,j}^{\text{YF2}}$$

Taking $\breve{Y}$ as the maximum value of Y the following three constraints help determine the value of product term $\chi_{i,j}^{\text{YF2}}$

$$\chi_{i,j}^{\text{YF2}} \leq \breve{Y} \cdot \chi_{i,j}^{\text{F2}}, \quad \chi_{i,j}^{\text{YF2}} \leq Y, \quad \chi_{i,j}^{\text{YF2}} \geq Y - \breve{Y} \cdot (1 - \chi_{i,j}^{\text{F2}})$$

The summation term $\sum T_{s_i}$ in equation (9), which represents the addition of T_s from all active users, can be written as following

$$\sum_{i=1}^{n} T_{s_i} = \sum_{i=1}^{Z} F_i \cdot T_{s_i}$$

Adopting this strategy equation (9) can be linearized as following

$$\sum_{i=1}^{n} F_i \cdot G = \sum_{j,k,l,m=1}^{Z} (A_{11} + A_{21} \cdot T_j) \cdot \chi_{j,k,l,m}^{\text{YF4}} + \sum_{j,k,l=1}^{Z} (A_{12} + A_{22} \cdot T_j) \cdot \chi_{j,k,l}^{\text{YF3}} + \sum_{j,k=1}^{Z} (A_{13} + A_{23} \cdot T_j) \cdot \chi_{j,k}^{\text{YF2}} + \sum_{j=1}^{Z} (A_{14} + A_{24} \cdot T_j) \cdot \chi_j^{\text{YF}} \quad (10)$$

It should be noted that equation (10) is valid for $n > 1$. If there is only one active user in the system then no medium contention would take place. In that particular case

$$E[D] = T_s + T_{\text{back-off}} = T_s + \frac{W-1}{2} \cdot \sigma = \widetilde{T} \quad (11)$$

Equation (10) and equation (11) can be combined by introducing another binary variable L, which is 1 if there is only one active user and 0 otherwise. The value of L is determined by following constraints

$$2 - L \cdot 10^9 \leq \sum_{j=1}^{Z} F_j \quad \text{and} \quad 1 + (1 - L) \cdot 10^9 \geq \sum_{j=1}^{Z} F_j$$

Finally, the linearized version of equation (9), which is valid for $n \geq 1$ is given as below

$$\sum_{i=1}^{n} F_i \cdot G = \sum_{j,k,l,m=1}^{Z} (A_{11} + A_{21} \cdot T_j) \cdot \chi_{j,k,l,m}^{\text{YF4}} + \sum_{j,k,l=1}^{Z} (A_{12} + A_{22} \cdot T_j) \cdot \chi_{j,k,l}^{\text{YF3}} + \sum_{j,k=1}^{Z} (A_{13} + A_{23} \cdot T_j) \cdot \chi_{j,k}^{\text{YF2}} + \sum_{j=1}^{Z} (A_{14} + A_{24} \cdot T_j) \cdot \chi_j^{\text{YP}} - \sum_{j=1}^{Z} \chi_j^{\text{YFL}} \cdot (A_{11} + A_{12} + A_{13} + A_{14} + (A_{21} + A_{22} + A_{23} + A_{24}) \cdot T_{s_j} - \widetilde{T_j}) \quad (12)$$

C. Applying the mathematical model in a simulation scenario

In the investigated scenario the LTE coverage is available in the whole area of user movement while WLAN coverage is limited in a circular area of 100 meter radius around a hotspot. This implies that the users always have LTE access available and WLAN coverage is only found in the vicinity of the hotspot (see Fig.4). During the resource assignment process, the flow management function classifies users into the following three categories (i) users with LTE access only and running VoIP or video applications (ii) users with LTE and WLAN access running any type of application (iii) users running FTP or HTTP applications with LTE access only. Users in the first category must be assigned the required minimum data rate through LTE as there is no other access available for them. Users in the second category are multi-homed users whose data rate will be decided by the aforementioned mathematical model. For users belonging to the third category, they must

 ISBN: 978-1-61208-229-5

get their traffic through the LTE path, however, it is not clear how much data rate should be allocated to them in order to achieve the optimized resource allocation objective. This issue is resolved by using the following work around: the users are assigned a $\widetilde{T_j}$ value greater than unity and they are put into the second category. The value of $\widetilde{T_j}$ greater than unity will refrain the LP solver to assign any data rate for these users over the WLAN path while the data rate for the LTE path will be decided based on the global objective of the optimized resource allocation. It is assumed here that each

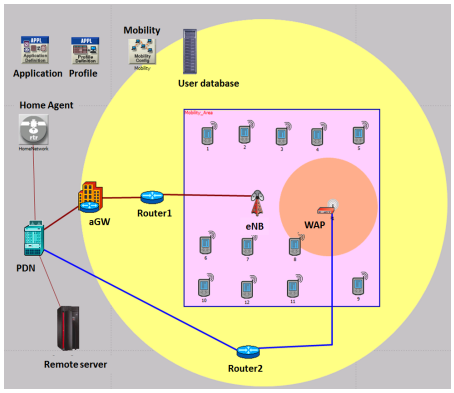


Fig. 4. Simulation scenario overview in the OPNET simulator. The large and small circular areas show the coverage of LTE and WLAN networks respectively. The user movement is restricted to the rectangular area.

user is running only one application. For a constant bit rate application, e.g., VoIP or video the minimum data rate is set equal to the maximum data rate in the model input parameters. For TCP based flows these two values can be set according to the network operator's policy. It should be noted that the problem has been formulated in a way that it guarantees the minimum data rate for all users and then assigns an additional data rate up to the maximum data rate while optimizing the spectral efficiency of the access networks.

The resource assignment process by the flow management function is carried out periodically in order to adapt to any changes in the user channel conditions. For this purpose, user channel condition parameters are obtained through cross layer information from the base stations of the two access technologies. According to this updated information, the mathematical model may reevaluate the solution considering the updated input parameters. As described earlier, the mathematical model is formulated using linear programming and solved using the C application programming interface (API) of ILOG CPLEX from IBM [7] which has been integrated inside the OPNET simulator by the authors. The output of this process consists of user data rates on each network path. These decided data rates are then conveyed to the users using fast LTE control plane signalling.

V. Simulation Results

This section shows the benefits of the proposed approach with the help of simulation results. For this purpose two scenarios are considered. In one scenario users do not make

TABLE I
Simulation Configurations

Parameter	Configurations
Total Number of PRBs	25 PRBs (5 MHz specturm)
Mobility model	Random Direction (RD) with 6 km/h
Number of users	3 VoIP, 2 Skype video call, 7 FTP uplink users
LTE Channel model	Macroscopic pathloss model , Correlated Slow Fading [13]
LTE MAC Scheduler	TDS: Optimized Service Aware, FDS: Iterative RR approach [12]
WLAN technology	802.11g, RTS-CTS enabled, coverage $\approx$ 100 m
VoIP traffic model	G.722.2 wideband codec, 23.05kbps data rate
Skype video model	MPEG-4 codec, 512kbps, 640x480 resolution, 30fps, play-out delay: 250 ms
FTP traffic model	FTP File size: constant 10 MByte continuous file uploads one after the other.
Simulation run time	10^3 seconds,10 seeds, 95% confidence interval

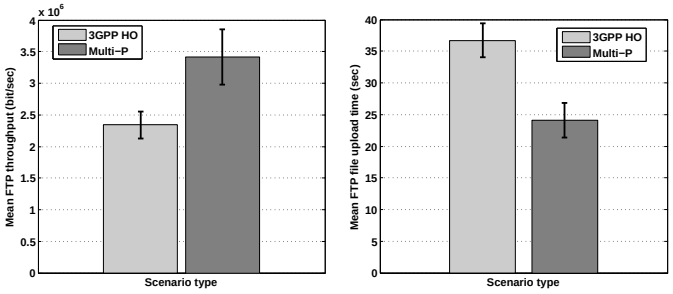
simultaneous use of LTE and WLAN access technologies. Instead the user traffic is completely handed over to WLAN as soon as the user is in the hotspot coverage, otherwise all traffic takes its path through the LTE access. This is the default policy for a multi-homed user according to the 3GPP specifications and therefore it will be referred to as "3GPP HO" case. Whereas, the second scenario extends the 3GPP architecture to support the simultaneous use of wireless interfaces, this will be referred to as "Multi-P". In this case, user traffic flows are distributed over the WLAN and the LTE access network. The traffic flow distribution policy is derived from the output of the optimization problem solved using linear programming. As a result, a user traffic flow is either sent over one network path with the least cost or it is split into two appropriately sized sub-flows each taking one network path to the destination.

Fig. 4 shows an overview of the simulation network model implemented in OPNET. The system is populated with 12 users generating a rich traffic mixture of: Voice over IP (VoIP), uplink File Transfer Protocol (FTP), and video conference (i.e., Skype video call). The users move within one LTE eNB cell, and within this cell one wireless access point (or hotspot) is present. The simulation configuration parameters are shown in Table I. Besides, in the "Multi-P" scenario the minimum data rate for FTP users is assigned as 200kbps while the maximum data rate limit is set to a very high value of 25Mbps.

In "3GPP HO" scenario, the users make handover between two access technologies without following make-before-break approach, i.e. the connection is broken from one network, and a new connection is established to the other one. Though MIPv6 keeps all IP layer connections alive through seamless handover, the user might lose data packets buffered at the lower protocol layers of the previously in use network interface. For example, LTE buffers the received IP packets at PDCP, RLC and MAC layers while WLAN keeps all the data buffered at MAC layer before transmission over the radio interface. Therefore, when making complete handover from one access technology to another, this buffered data is discarded and have to be recovered by upper layers through retransmissions. This behavior leads to applications performance degradation for both TCP and UDP based applications.

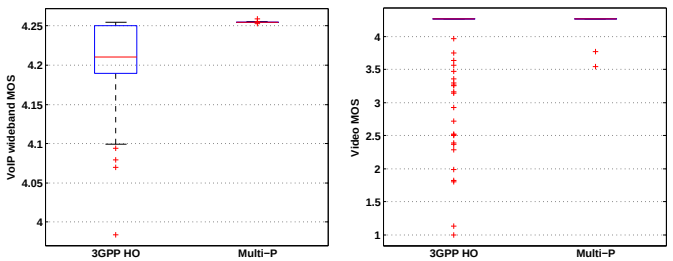
 ISBN: 978-1-61208-229-5

The "Multi-P" scenario the users are allowed to use the WLAN access when it is in the coverage, and can still keep the LTE connection alive and use it at the same time. In the coverage of WLAN access, the flow management client function sends user traffic on WLAN link only when user PHY mode is 9Mbps or higher. This is because when a user enters in 6Mbps mode it implies that the user is almost at edge of coverage which is a strong indication that loss of WLAN link is imminent. Hence, no new traffic data is scheduled for WLAN link which gives user a chance to transmit already buffered data to the access point before the loss of link happens. Moreover, "Multi-P" approaches, in contrast to "3GPP HO" scenario, keep buffered data at the minimal required level through the use of network path capacity estimations.



(a) Mean user FTP uplink throughput (b) Mean user file upload time

Fig. 5. User evaluation for non-real time application



(a) VoIP call quality (b) Video call quality

Fig. 6. User evaluation for real time applications

Fig. 5 shows the performance of FTP uplink application for the users. It can be seen that the "Multi-P" algorithm achieves the best results for FTP uplink user traffic. The figure shows that "Multi-P" provides 32% higher user FTP throughput than "3GPP HO" scenario. The higher throughput helps users finish file upload faster. The gain in FTP throughput is mainly coming from the proper management of network resources where users with good channel conditions are assigned more resources in LTE network. Similarly in WLAN network users with the best channel condition are allowed to transmit.

Fig. 6 shows the boxplot of user Mean Opinion Score (MOS) values of VoIP and video services. The MOS values of the wideband VoIP codec and video codec are computed using the modified E-model and Evalvid toolkit as described in [4] and [3], respectively. The figure shows that VoIP and Video users in "Multi-P" scenario mostly achieve the best possible MOS value. The users in "3GPP HO" scenario often suffer from quality loss when making handover as well as when transmitting over QoS unaware WLAN network. This shows that proper management of network resources as performed by flow management function not only improves the network capacity but also enhances user QoE.

VI. Conclusion

This work highlighted the importance of multi-homing support in integrated heterogeneous wireless networks of 3GPP and non-3GPP access technologies. The existing 3GPP specifications for the integration of two types of the access technologies (i.e. 4G LTE and WLAN) are extended following IETF standards to realize multi-homing support for the users. This work mainly focuses on the problem of optimum resource utilization in such a heterogeneous network where the users and network operators can take advantage of multi-homing support. The problem of optimum network resource allocation is mathematically modeled using the linear programming technique. The mathematical model is then integrated in the network simulator to decide the network resource allocation for multi-homed users during the simulation. With help of simulation results it is shown that the proposed scheme of resource allocation brings twofold gain when compared to the 3GPP proposal. On the one hand it significantly improves the network capacity and on the other hand it fulfills the user application QoS demands which otherwise cannot be satisfied from QoS unaware non-3GPP access technologies.

References

[1] 3GPP Technical Report TS 36.213, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures", v10.2.0, Jun. 2011.

[2] Toseef, U., Zaki, Y., Timm-Giel, A., and Görg, C., "Development of Simulation Environment for Multi-homed Devices in Integrated 3GPP and non-3GPP Networks", The 10th ACM International Symposium on Mobility Management and Wireless Access in Paphos, Cyprus, Oct. 2012.

[3] J.Klaue, B. Rathke, and A. Wolisz, "EvalVid - A Framework for Video Transmission and Quality Evaluation", In Proc. of the 13th International Conference on Modeling Techniques and Tools for Computer Performance Evaluation, pp. 255-272, Illinois, USA, Sept. 2003.

[4] Toseef, U., Li, M.,Balazs, A., Li,X., Timm-Giel A., and Görg, C., "Investigating the Impacts of IP Transport Impairments on VoIP service in LTE Networks," in 16. VDE/ITG Fachtagung Mobilkommunikation, Osnabrück, Germany, May 18-19, 2011.

[5] 3GPP Technical Report TS 23.402, "Architecture enhancements for non-3GPP accesses, 3rd Generation Partnership Project", v10.6.0, Dec. 2011.

[6] OPNET, http://www.opnet.com, as accessed in Oct. 2012.

[7] CPLEX, http://www.ibm.com/software/, as accessed in Oct. 2012.

[8] Yasir Zaki, "Long Term Evolution (LTE) Model Development Within OPNET Simulation Environment", OPNET Work Dec. 2011, Washington.

[9] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," IEEE Journal on Selected Areas in Communications, Vol. 18, No. 3, pp. 535-547, Mar. 2000.

[10] R. Litjens, F. Roijers, J. L. van den Berg, R. J. Boucherie, and M. Fleuren, "Performance Analysis of wireless LANs: an Integrated Packet/Flow Level Approach", ITC Conference, Berlin, Aug. 2003.

[11] P. Chatzimisios, V. Vitsas, and A.C. Boucouvalas, "Throughput and Delay Analysis of IEEE 802.11 Protocol", in Proceedings of the 5th IEEE International Workshop on Networked Appliances (IWNA 2002), pp. 168-174, Liverpool, UK, 30-31 Oct. 2002.

[12] S. N. K. Marwat, "Performance Evaluation of Bandwidth and QoS Aware LTE Uplink Scheduler",10th International Conference on Wired/Wireless Internet Communications, Santorini, Greece, Jun. 2012.

[13] 3GPP Technical Report TS 25.814, "Physical layer aspects for E-UTRA", 3rd Generation Partnership Project, v.7.1.0, Sept. 2006.

 ISBN: 978-1-61208-229-5

Methods and Issues in Detecting Pedestrian Flows on a Mobile Adhoc Network

Ryo Nishide, Hideyuki Takada
Faculty of Information Science and Engineering
Ritsumeikan University
Kusatsu, Japan
nishider@fc.ritsumei.ac.jp, htakada@cs.ritsumei.ac.jp

***Abstract*—Due to the development of mobile technology and adhoc communication, researches to extract social contexts including the movements and density of pedestrians have also emerged in recent years. This study attempts to explore methods to extract pedestrian flows in a distributive manner, deploying Bluetooth detection logs. Bluetooth devices are widely installed in mobile equipments which pedestrians carry with them in daily life. The results of experiments have revealed that detection logs implicitly record traces of surrounding pedestrian flows, which might provide possibilities to analyze and distinguish pedestrian flow patterns based on situations. Moreover, the paper has discussed the related issues on network construction including methods for interpolating missing detections.**

Keywords-Distributive Database; Bluetooth; Social Context; Pedestrian Flows; Mobile Devices; Adhoc Network.

I. INTRODUCTION

According to the increase of urban population and the expansion of social activities, we cannot avoid sharing the same public spaces with other people when traveling as well as in daily life. In any occasion, it will be one of the major concerns for people whether the area is crowded or less-crowded, and sometimes, it is necessary to know what is actually going on in such places, including the changing flow of pedestrians. On the other hand, many location-based services have appeared on market owing to the enhancement of computational ability and wireless communication technology, such as WiFi [1] and Bluetooth [2], and GPS technology [3] deployed in mobile devices. These advancements have paved way to explore methods for detecting pedestrian flows or social contexts using high performance mobile devices [4].

This research employs methods to extract the density and flows of pedestrians using the Bluetooth detection logs, while considering the data management scheme on a mobile adhoc network. This adhoc network can be generated from connection between mobile devices to work as a distributive database, which can be managed and updated the detection log data, or modified the log data by accessing to geometrically adjacent devices to check for missing detections. The policy of this work is to avoid initial preparations, such as installing a large number of expensive immovable sensors and high performance computational equipments in real space, in order to minimize cost, time and effort. In this research, we focus on the attempt to extract pedestrian flows in real world, while the specific services to utilize the detection results are left for future works.

We attempt to grasp social contexts such as changes of pedestrian flows and density by detecting the surrounding electronic equipments. Recent handheld electronic equipments like cell phone, smart phone, PDA, and laptop are installed with wireless devices such as WiFi and Bluetooth, which pedestrians carry with them in their daily lives. If these devices surrounding the user are detected and logged continuously, it may be possible to detect not only the density of crowd, but also the changes of pedestrian movements.

We have conducted a preliminary investigation to examine the statistics of detectable types of terminal (mobile phone, PC, etc.) at various places [5]. Comparing two wireless technologies, WiFi and Bluetooth, WiFi was detected from many types of electronic equipments either carried by pedestrians or fixed in the environment. Therefore, WiFi seems difficult to discriminate the types of equipments, whether they are carried by the pedestrians or not. On the other hand, by the investigation of Bluetooth signals, most of the detected Bluetooth radios were from mobile devices. In this paper, we focus on Bluetooth devices installed in equipments to be carried by users in order to examine the flows and movements of pedestrians.

To begin with, Section II provides the method of detecting pedestrian flows. The detection results obtained from actual experiment have been examined in Section III. Through the experiment, several devices were not detected within the scanning interval. Therefore, distributive and autonomous method to interpolate missing detection is discussed in Section IV and V in order to enhance further analysis of log data.

II. DETECTION OF PEDESTRIAN FLOWS

We avoid extracting the personal information of pedestrians, such as locations or user's name, since collection of such information might violate the privacy issue of pedestrians. Instead, we examine the detection patterns (e.g., numbers and changes of simultaneous or continuous detections) of devices carried by pedestrians surrounding the user. Fig. 1 shows an example of the pedestrians' Bluetooth devices which have entered the reachable communication range of

 ISBN: 978-1-61208-229-5

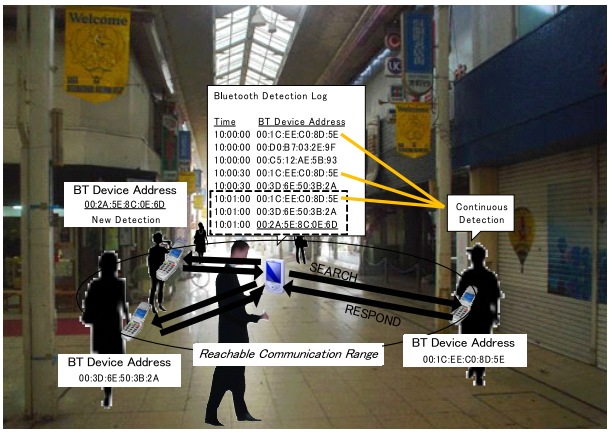


Figure 1. Detection of Pedestrian Flows

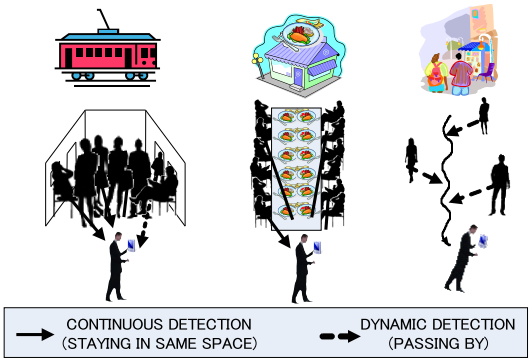


Figure 2. Detections in Different Situations

the user who is conducting the experiment. User's device continuously sends inquiry to search for the surrounding pedestrians' devices, and logs the time and Bluetooth Device Address (BDA) of devices which have responded to user's inquiry. BDA is a unique ID (MAC address) assigned to each bluetooth device during manufacture process for the purpose to identify each device. From the examination of logs, different types of detections can be verified, such as continuously detected, newly detected, undetected or disappeared, and so on, which might be the key to determine the flows of dynamic pedestrians in real world. Since the detection patterns differ depending upon the situations of the surrounding pedestrians (Fig. 2), it might be possible to assume the social contexts or trends and changes of surrounding situations by analyzing such detection patterns.

III. Verification of Detection Patterns

We have done several investigations to observe surrounding Bluetooth devices in various situations. To collect data, we used HP iPAQ 112 Classic Handheld PDA, which has been set to record BDA with a timeout interval of 6 seconds after sending inquiry signal for every 30 seconds cycle.

Four different cases have been examined in this paper, namely strolling in town, transporting by train, attending the conference, and taking lunch at a cafeteria. The results of examination of detection logs are summarized in Fig. 3. The upper diagram of Fig. 3 shows the detection pattern of Bluetooth devices, with the time-line expressed on the horizontal-axis, and the device ID assigned in chronological order of the incoming BDA on the vertical-axis. The mobile phones are colored in red, and PCs and devices other than mobile phones in green, and unidentified devices in blue. The lower diagram of Fig. 3 shows the number of detected devices, with the time-line expressed on the horizontal-axis, and the quantity of BDA on the vertical-axis.

(a) Strolling in Town: Fig. 3(a) shows the changes of multiple detection logs encountered while strolling in town. The number of BDAs is not constant as the number of passers-by is always changing. Even if the pedestrians are walking in the same direction, their devices disappeared occasionally probably because their directions coincided only for a while or their walking speed was different. On the other hand, the same BDA was continuously identified in some places while the examiner was dropping in stores.

(b) Transporting by Train: Fig. 3(b) shows the detection in the train during rush hours. From the log, we can verify such situations as: (i) devices were continuously detected from passengers in the same car; (ii) many incoming and outgoing devices were detected when changing the trains; and (iii) a large number of people got on/off the train at major stations. The passenger's devices can be constantly detected while the train is moving. However, due to the limited size and shape of the car, the detection has been low even in rush hours.

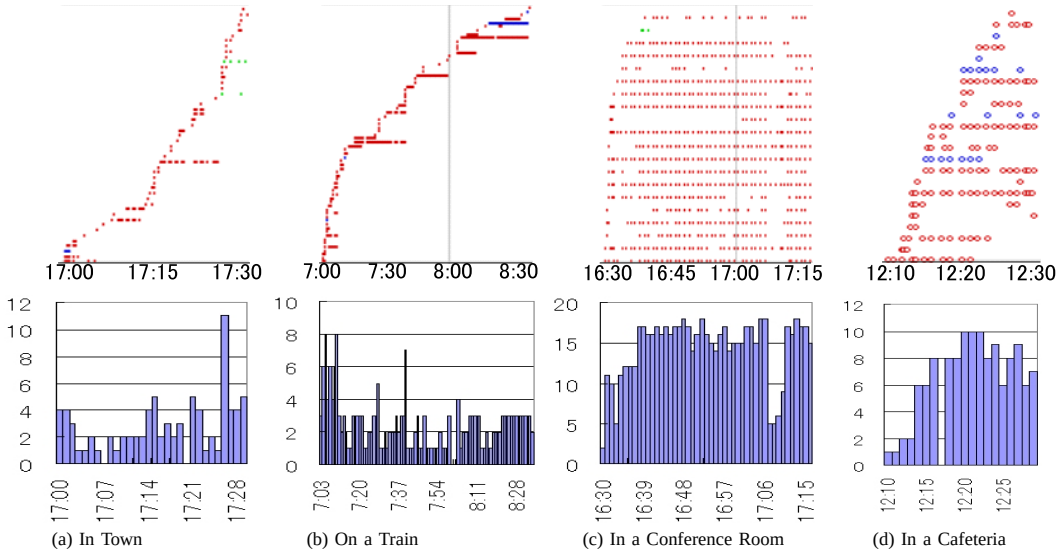
(c) Attending the Conference: Fig. 3(c) shows that many BDAs were detected continuously in the same room. As most of the participants were staying in the room during the conference, the number of BDAs was almost constant (14 to 18 devices), except the time for coffee break. As the room was wide enough to hold many people, the quantity of detection has been kept high.

(d) Taking lunch at a Cafeteria: Fig. 3(d) shows that many devices have been detected during lunch time, as customers enter, take lunch and leave the cafeteria one after another. Some devices are detected continuously with long duration, and others are divided into several times with short duration, because two types of situations are mixed together: people sitting and eating lunch, and people walking around to look for seats or friends.

These results show possibilities that the pedestrian flow can be assumed by analyzing the detection logs as follows:

- **The number of BDA detection log:** crowdedness of people (requiring reference to the scale of space)
- **Time length of BDA detection:** people staying in same space or duration of the event
- **Appearance/Disappearance in BDA detection:** people staying, entering, leaving, or passing by

 ISBN: 978-1-61208-229-5



(a) In Town (b) On a Train (c) In a Conference Room (d) In a Cafeteria

Figure 3. Detection Pattern of BDA (upper), Detected Number of BDA (lower)

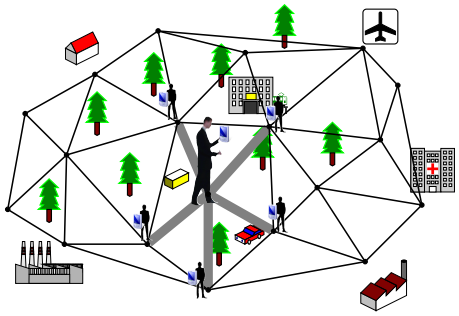


Figure 4. Delaunay Network with Mobile Devices

The detection logs show that there are several undetected devices even among those staying in the same space. Therefore, a method to interpolate the missing detection is also explored in following sections.

IV. Scheme for Distributive Data Management

Another important issue of concern is the management scheme of pedestrian flow data obtained from each mobile device. It is not efficient to collect and manage the entire data sent from mobile devices on a server. Therefore, a mechanism is necessary to manage data and perform computation between mobile devices cooperatively.

In this paper, we apply the method proposed in the past works [6] to generate P2P Delaunay network, which is a geometry-based P2P network whose topology is defined by the geometric adjacency of mobile devices as illustrated in Fig. 4. It has the features as following: (i) each device connects to a close-by devices based on its geographical distance; (ii) the degree of connection for each device is low (approximately six); (iii) the network can correspond with join/leave of device only affecting the surrounding devices to reconstruct and update the connection; and (iv) the data is reachable to distant device through multi-hop communication.

We assume that each mobile device only has the location information of other devices, but not the knowledge of how the other devices are connected. Thus, each mobile device must choose the appropriate mobile devices to connect, referring to their location information to generate a P2P Delaunay Network. The detail algorithm for generating and maintaining connections are discussed in the past works [6]. Delaunay Network can be used not only to generate or maintain connections with adjacent nodes on a plane, but also to perform collaborative computation with adjacent nodes described in the following section.

V. Interpolation of Missing Detection

There are false-negative cases that some devices within the communication range may not be detected. To deal with such problems, we consider methods to check the detection logs of adjacent nodes on Delaunay network, and interpolate the BDA data which is definitely within the communication range of Bluetooth device. Initially, each node sends a copy of its own detection logs to adjacent nodes, and receives their copy of detection logs. Then, it extracts the BDA data which is not detected from its device, but detected from other adjacent devices. These BDA data will be the target data to perform interpolation, and the location of these adjacent nodes will be the criterion to determine whether or not to perform interpolation.

We validate only the BDA data owned by more than three adjacent nodes to perform interpolation. That is, a polygon is drawn using the location of adjacent nodes with the target BDA data as vertices. If the location of its own node is within the polygon, then the target BDA data shall

 ISBN: 978-1-61208-229-5

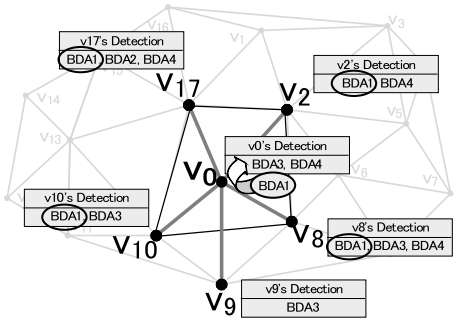


Figure 5. Interpolation of BDA Data (BDA1)

be the one to be interpolated. We chose polygonal shape to determine the interpolation, because it is obvious that the entire polygonal region is covered from the communication range of Bluetooth device. The purpose of this interpolation method is to deal with missing detection, and the deformation of communication range caused by walls, buildings, and obstacles are beyond our focus.

Fig. 5 shows the interpolation process using the Delaunay Network. Node v_0 has five adjacent neighbor nodes, namely v_1, v_2, v_3, v_4, v_5, and has the copy of their BDA detection logs. Among the BDA on detection logs, BDA1 is the only one that v_0 does not have, but more than three adjacent nodes (v_1, v_2, v_4, v_5) have. Using these nodes as vertices, a polygon is drawn starting from the upper node in clockwise direction. Finally, BDA1 can be determined to be included in v_0's detection data, as it is allocated within the polygon.

VI. Related Works

Several researches have emerged in the attempt to extract social contexts, owing to the development of mobile equipment and adhoc communication.

O' Neill et al. [7] and Nicolai et al. [8] examined the correlation between Bluetooth detecting and pedestrian movement by deploying stationary Bluetooth sensors in the environment and analyzing the logs. Eagle et al. [9] has shown methods to analyze social patterns of user's activity in a daily routine. These works show that Bluetooth scanning and analysis of detection logs have possibility to extract the flow of pedestrians, however, not every Bluetooth device can be guaranteed to be detected depending upon the performance of the device and situation of the space.

To cope with such problem, Kim et al. [10] examined the detection pattern of Bluetooth device logs, and employed clustering algorithm and Gaussian blur to remove noises caused by inquiry fault of undetected Bluetooth devices. They inferred the transition time of events from multiple device detections. Weppner et al. [11] estimated crowd density through collaboration with multiple devices to improve the accuracy of detections. Users were assigned to carry multiple devices to perform Bluetooth scanning together, which might be troublesome for users.

VII. Conclusion

We have shown possibilities to extract pedestrian flows by detecting the surrounding Bluetooth devices, and proposed to apply distributive methods to generate mobile adhoc network and manage detection data on the network cooperatively. For future works, we plan to perform detailed analysis on Bluetooth device logs, examine the applicability with other sensory data, and provide location-based application using pedestrian flows as social contexts. On the other hand, we plan to continue further study on Delaunay networks, exploring efficient ways to manage social contexts data and log files, while evaluating proposed methods to interpolate missing data caused by inquiry faults.

References

[1] Wi-Fi Alliance, http://www.wi-fi.org (Retrieved August 23, 2012)

[2] Bluetooth Technology, http://www.bluetooth.com (Retrieved August 23, 2012)

[3] B. Parkinson and J. Spilker Jr., Global Positioning System: Theory and Applications, American Institute of Aeronautics and Astronautics, 1996. ISBN 978-1-56347-106-3.

[4] P. Lukowicz, A. Pentland, and A. Ferscha, "From Context Awareness to Socially Aware Computing," IEEE Pervasive Computing, 11 (1), pp. 32–41, 2012.

[5] R. Nishide, T. Ushikoshi, S. Nakamura, and Y. Kono, "Detecting Social Contexts from Bluetooth Device Logs", Supplement Proc. Ubiquitous Computing (UbiComp), pp. 228–230, 2009.

[6] M. Ohnishi, R. Nishide, and S. Ueshima, "Incremental Construction of Delaunay Overlaid Network for Virtual Collaborative Space", 3-rd Proc. Conf. on Creating, Connecting and Collaborating through Computing (C5'05), IEEE CS Press, pp. 77–84, 2005.

[7] E. O'Neill, V. Kostakos, T. Kindberg, A.F. Schiek, A. Penn, D. Fraser and T. Jones. "Instrumenting the city: Developing methods for observing and understanding the digital cityscape", UbiComp, pp. 315–332, 2006.

[8] T. Nicolai and H. Kenn. "About the relationship between people and discoverable bluetooth devices in urban environments", 4th int'l conf on mobile technology, applications, and systems, pp. 72–78, 2007.

[9] N. Eagle and A. Pentland. "Reality mining: sensing complex social systems", Personal and Ubiquitous Computing, 10, pp. 255–268, 2006.

[10] D. Kim and D.-K. Cho, BlueSense: Detecting individuals, locations, and regular activities from Bluetooth Signals, http://www.cs.ucla.edu/∼dhjkim/files/pdf/cs219_BlueSense.pdf (Retrieved August 23, 2012)

[11] J. Weppner and P. Lukowicz, "Collaborative Crowd Density Estimation with Mobile Phones," In Proc. ACM Workshop on Sensing Applications on Mobile Phones at ACM SenSys, pp.26–30, 2011.

 ISBN: 978-1-61208-229-5

On Demonstrating Spectrum Selection Functionality for Opportunistic Networks

Alessandro Raschellà, Anna Umbert, Jordi Pérez-Romero, Oriol Sallent
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
e-mail: [alessandror,annau,jorperez,sallent]@tsc.upc.edu

Abstract— **This paper presents a testbed platform to demonstrate and validate spectrum opportunity identification and spectrum selection functionalities in Opportunistic Networks (ONs). The hardware component of the testbed is based on reconfigurable devices able to transmit and receive data at different operating frequencies, which are dynamically configured. The software component has been developed to perform the creation and maintenance of ON radio links, including spectrum opportunity identification and selection decision making as well as all the necessary signaling to support the ON operation. Therefore, the presented platform provides a powerful tool for testing different algorithms in real operational radio environments under various interference conditions, thus enabling to gain deeper insight into the performance of algorithmic solutions, beyond the purely theoretical analyses based on models and/or simulations. Results presented in the paper validate the implementation conducted at the laboratory and illustrate the reconfigurability capabilities of the ON links under different conditions.**

Keywords-Opportunistic Networks; Spectrum Selection; Spectrum Opportunity Identification; Testbed.

I. INTRODUCTION

It has been stated that the Internet has been successful because of its flexibility, its accessibility via different physical media, and for its simple support of many different types of applications and data types. Initially, wired access was dominant, while the set of applications was limited mainly to file-transfer, e-mail, media streaming and client-server based web services/applications. In many positions on the Future Internet (FI), wireless access is expected to prevail, while at the same time there is growing interest for more application (deployment) areas; thus, the FI is penetrating and covering almost every facet of our lives. For instance, increasingly modern information and communication services are built around social network concepts that require smart personal devices, and this makes it even more imperative to meet the need to offer appropriate connectivity everywhere where media or data flows need to be provided. Diversified applications/services can be accessed at any time of a day, can be requested from all types of locations/environments (e.g., home, public, work, urban, rural, etc.) or by all types of communication end-points (e.g., machines, humans acting in different roles, namely in-work or private life), and can involve various information flows (voice, audio, data, images, video) and communication types (uni-cast, multicast, broadcast, peer-to-peer). In contrast to today's Internet, for the FI, it can be safely assumed that the "best effort" delivery model will not hold. Certain applications, services and content will have to be delivered under Quality of Service (QoS) levels, or at least guaranteeing a certain Quality of Experience (QoE).

Such hard requirements will set the networks under an enormous stress for resources (bandwidth, storage processing required) in both core and access parts. Traditionally, the need for more resources has been addressed through worst-case (peak-hour) based planning. This has lead to over-provisioning of resources in non-peak times. Keeping in mind that wireless resources are "expensive" (in the sense of "limited" or "scarce"), this over-provisioning will have to be tackled. In this respect, a range of solutions have been applied. For instance, many operators are aggressively adding WiFi access points and femtocell nodes to their network, in order to offload large portions of the traffic from the wide area networks of their infrastructure. However, as user behavior changes and user expectations increase, so do the resource requirements that are posed onto the communication networks. These continuously increasing requirements motivate the quest for further efficiency in resource provisioning.

Opportunistic Networks (ONs) are considered as an innovative solution to satisfy the demand for applications/services and respective resources, through increased efficiency in resource provisioning and utilization [1]. ONs are temporary, localised network segments created under certain circumstances. In this vision, ONs are always governed by the radio access network (RAN) operator (which provides the resources, the policies, the knowledge, etc.) so they can be considered as coordinated extensions of the infrastructure. ONs comprise both infrastructure nodes and infrastructure-less devices. The aim for a RAN operator to use ONs is to improve the performance of the infrastructure network, but also (and perhaps via a third party) to provide a new span of localised or closed-group services. Further on, the introduction of cognitive techniques for the management of the ONs will lead to robustness and to capitalize the learning capabilities intrinsic to cognitive systems.

ONs have been investigated to provide efficient solutions for a wide range of possible scenarios and use cases [2], such as: (1) "Opportunistic coverage extension", which describes a situation in which a device cannot connect to the operator's infrastructure, due to lack of coverage or a mismatch in the radio access technologies. The proposed solution includes an additional connected user that, by creating an opportunistic network, establishes a link between the initial device and the infrastructure, and acts as a data relay for this link. (2) "Opportunistic capacity extension", which depicts a situation

 ISBN: 978-1-61208-229-5

in which a device cannot access the operator infrastructure due to the congestion of the available resources at the serving access node. The solution proposes the redirection of the access route through an ON that avoids the congested network segment. (3) "Infrastructure supported opportunistic ad-hoc networking", which shows the creation of a localised, infrastructureless ON among several devices for a specific purpose (peer-to-peer communications, home networking, location-based services, etc.). Infrastructure governs the ON creation, benefits from the local traffic offloading and develops new opportunities for service provisioning.

A common technical challenge in the different scenarios and ON use cases is to decide the proper spectrum to be used for the transmission of data and control flows in any communication link in accordance with the requirements for this link depending on the applications to be supported. This functionality is referred to as spectrum selection and it envisages a dynamic and flexible use of the available spectrum that ensures an efficient usage of this resource. The spectrum management process should be divided in two differentiated steps. First, the spectrum opportunity identification will be in charge of finding out the set of possible frequency bands that are available for the link. Second, and based on the results of the previous step, the spectrum selection will decide the most adequate band for the communication. Spectrum opportunity identification and spectrum selection functionalities have been a topic of research in different studies. For instance, [3]-[5] proposed energy detection as a means to identify spectrum opportunities, while [6]-[10] present different algorithms and protocols for assigning spectrum in cognitive radio networks.

In this context, this paper describes the testbed implementation platform that has been developed for demonstrating and validating the spectrum selection functionality in ONs. It is built based on reconfigurable devices able to operate in different frequencies dynamically configured. This allows establishing and monitoring ON radio links, and reconfiguring them based on the changes in the current spectrum conditions. In this way, the testbed provides a practical insight for testing different algorithms in real environments, going beyond the purely theoretical analyses based on models and/or simulations.

The rest of the paper is structured as follows. In Section II, the ON life cycle and functional architecture for ON management are presented. Then Section III presents the algorithmic solutions for spectrum selection considered in the testbed, and Section IV provides the testbed implementation. Section V presents some results and Section VI summarises the conclusions and next steps.

II. OPPORTUNISTIC NETWORKS: LIFE CYCLE AND FUNCTIONAL ARCHITECTURE

The life cycle of an ON comprises the following phases: (1) Suitability determination, where the convenience of setting up a new ON is assessed according to the triggering situation, previous knowledge, policies, profiles, etc., (2) Creation, which includes the selection of the optimal, feasible configuration for the new ON (selection of the participant nodes, the spectrum and the routing pattern), (3) Maintenance, which involves monitoring and controlling the QoS of the data flows involved in the ON and performing the appropriate corrective actions when needed, and (4) Termination, when the motivations for the creation of the ON disappear or the ON can no longer provide the required QoS and, therefore, mechanisms should be provided to handle handovers and to keep applications alive if possible.

Spectrum selection is involved in all the management stages in the ON life cycle. During suitability determination, which is the result of a rough feasibility analysis in order to keep complexity moderate, there is the need to introduce mechanisms leading to the identification of spectrum opportunities that ensure that the resulting interference conditions in the possible future ON will result acceptable. The suitability stage will provide one or several possible configurations for an ON, whose feasibility and potential gains have been roughly estimated. Then, during the creation a detailed analysis (thus probably requiring additional context awareness and/or more accurate estimations related to diverse aspects of the radio environment) will be conducted and the spectrum to be assigned will be decided.

ON reconfiguration capabilities will provide the necessary adaptability to changing conditions. This stage comprises monitoring (i.e., dynamically acquire all the relevant information that may influence decision making processes around the ON such as relevant changes in the radio spectrum occupancy/interference conditions) and reconfiguration decisions. Reconfiguration decisions will be supported by other functionalities like discovery procedures for the identification of new nodes, identification of spectrum opportunities, etc.

Based on the functional architecture proposed in [11] by the European Telecommunications Standardization Institute (ETSI) for Reconfigurable Radio Systems (RRSs) an extension was proposed in [12] to deal with ON management. ON management features are attributed to an entity implemented in terminal/infrastructure. Fig. 1 depicts an example and simplified view of such management entity at the infrastructure, highlighting (1) the decision-making processes associated to the different ON stages, (2) the control mechanisms that will lead to execute the decisions taken, (3) the knowledge management module to exploit cognitive features, (4) the context awareness to provide the necessary inputs about the radio environment conditions to the decision making algorithms, and (5) the Dynamic Spectrum Management (DSM) that provides the spectrum availability conditions and related constraints to guide the spectrum selection decision making.

III. SPECTRUM SELECTION: ALGORITHMIC SOLUTIONS

From an algorithmic perspective, the problem considered in the testbed presented in this paper is the selection of the spectrum to be assigned to a set of radio links between a pair of terminals and/or infrastructure nodes. The purpose of each radio link is to support a given application with certain bit rate requirements. The spectrum selection is carried out in the decision making entity and is supported by the spectrum opportunity identification residing in the DSM module.

 ISBN: 978-1-61208-229-5

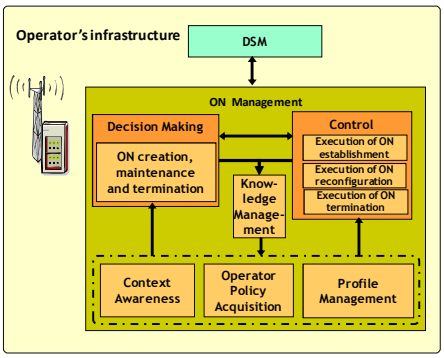


Figure 1. ON management at the infrastructure side

A. *Spectrum Opportunity Identification algorithm*

The spectrum opportunity identification algorithm executes two different procedures: the measurement procedure and the spectrum block formation.

In the measurement procedure, the total analysed band is subdivided into N smaller portions of equal band Δf. The measurement algorithm performs an energy detection sensing (during a period of time Δt) for each Δf portion until measuring the total band, starting from frequency *F_min_band*. This measurement is repeated *Num_Meas* times. Then, based on the multiple measurements carried out, the Spectrum Opportunity Index (SOI) is obtained for each portion, defined as the fraction of measurements in which this portion has been detected as available. The power threshold to decide if a portion is free is set based on [13].

In the spectrum block formation procedure, the consecutive spectrum blocks with SOI above a certain threshold are grouped in blocks. Each block is constituted by a maximum of P_{max} portions. For each block, the algorithm returns the 2-tuple $SB_k=\{f_k,BW_k\}$ where f_k is the central frequency of the block and BW_k the bandwidth.

B. *Spectrum selection algorithm*

The spectrum selection algorithm uses as input the set of available spectrum pools resulting from the spectrum opportunity identification, together with the characteristics of each pool in terms of available bit rate based on radio considerations. The algorithm output is the list of spectrum assignments to each of the existing links. The algorithm presented in [14] is considered for the implementation in the testbed. It makes use of the fittingness factor concept as a metric to capture how suitable a specific spectrum pool is for a specific radio link. The algorithm is based on estimating the fittingness factor for each link and available spectrum block based on a knowledge database that is maintained with different fittingness factor statistics.

IV. TESTBED IMPLEMENTATION

In this section, the testbed implementation is provided; in details, the hardware and software components and the testbed architecture are illustrated.

A. *Hardware component: basic building block*

The testbed demonstrator is built on the basis of Universal Software Radio Peripheral (USRP) boards. Each USRP integrated board incorporates AD/DA Converters (ADCs/DACs), a Radio Frequency (RF) front end, and a Field Programmable Gate Array (FPGA) which does some pre-processing of the input signal [15]. A typical setup of the USRP board consists of one mother board and up to four daughter boards. On the mother board, there are four slots, where up to 2 RX and 2 TX daughter boards can be plugged in. The daughter boards are used to hold the radio frequency receiver and the radio frequency transmitter. There are 4 high-speed 12-bit ADCs and 4 high-speed 14-bit DACs. All the ADCs and DACs are connected to the FPGA that performs high bandwidth math, such as interpolation and decimation. The DACs clock frequency is 128 Ms/s, while ADCs work at 64 Ms/s to digitize the received signal. A USB controller sends the digital signal samples to a PC in I/Q complex data format (4 bytes per complex sample), resulting in a maximum rate of 8 Ms/s. Consequently, the FPGA has to perform filtering and digital down-conversion (decimation) to adapt the incoming data rate to the USB 2.0 and PC computing capabilities. The maximum RF bandwidth that can be handled is thus 8 MHz.

There exist different kinds of daughter boards that allow a very high USRP reconfigurability and working at several frequency bands. The daughter boards integrated in the USRP motherboard of this testbed are *XCVR2450 Transceivers*. They work in the frequency ranges 2.4 - 2.5 GHz and 4.9 - 5.9 GHz.

B. *Software Component*

Identification of spectrum opportunities is performed by both a hardware platform (i.e., USRP) and a software component implemented with GNU Radio toolkit. It is a software for learning about, building and deploying software radios [16]. GNU Radio is free and open source. It provides a library of signal processing blocks and the glue to tie it all together. In GNU Radio, the programmer builds a radio by creating a graph (as in graph theory) where the vertices are signal processing blocks and the edges represent the data flow between them. All the signal processing blocks are written in C++ and Python is used to create a network or graphs and glue these blocks together. GNU Radio has been used to develop the modules that implement the algorithms described in Section III and to enable the data and control communication between USRP transceivers.

C. *Testbed architecture*

The objective of this testbed is to show the behaviour of the spectrum opportunity identification and spectrum selection procedures in an ON. For that purpose, a scenario is considered where two devices need to communicate through an ON link controlled by the infrastructure, as graphically illustrated in the upper part of Fig. 2. Both spectrum opportunity identification and spectrum selection functionalities reside in the infrastructure node. The result of executing these functions, with the specific frequency block assigned for the ON link between the two terminals is

 ISBN: 978-1-61208-229-5

notified using a Cognitive Control Channel [17].

The testbed implementation of the infrastructure node and the terminals by means of USRP transceivers is shown in the lower part of Fig. 2. USRP#1 implements the infrastructure and the associated spectrum identification and selection functionalities, while USRP#2 and USRP#3 are the terminals exchanging data. ISM 2.4 GHz band is used for the demonstration, detecting the available spectrum opportunities and allocating a portion of this band for the communication between terminals.

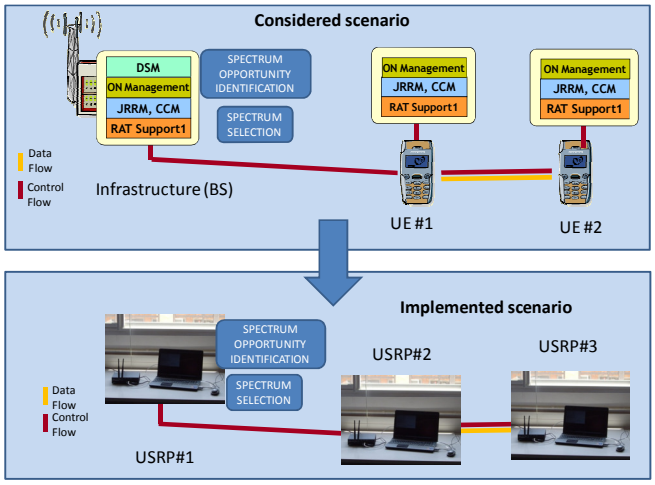


Figure 2. Scenario considered in the demonstration and corresponding implementation by means of USRP

D. *Signalling procedures*

Since the target of the demonstration is the spectrum opportunity identification and spectrum selection, the demonstration implements only the ON creation and ON maintenance stages of the ON life cycle. It is assumed that the decision to create an ON among the two devices has been previously made in the ON suitability phase.

The cognitive control channel signaling is implemented with the Control Channel for the Cooperation of the Cognitive Management System (C4MS) protocol using the implementation option based on IEEE 802.21 "Media-Independent Handover (MIH) Services" [18]. The implemented procedure for the ON creation is shown in Fig. 3. The steps of the procedure are explained below.

1. A MIH_C4MS_ONN.request message is sent from UE#1 to the infrastructure (Base Station - BS) to start the ON-Negotiation (ONN) procedure intended to obtain a valid configuration of the radio link. The message indicates the terminals involved and the QoS requirements that the link is expected to support, in terms of required bit rate.

2. The infrastructure sends a MIH_C4MS_ONN.request to UE#2 informing it about the intention to establish a direct radio link with UE#1 and allowing it to join the negotiation process for the derivation of the radio link configuration.

3. UE#2 replies to the BS with a MIH_C4MS_ONN.response message, notifying its acceptance for the establishment of the link.

4. The ON management entity in the infrastructure inquires the DSM entity to determine spectrum availability for the link. The spectrum opportunity identification algorithm is executed.

5. DSM reply provides the available spectrum blocks, and the spectrum selection algorithm is executed to decide the spectrum block to be allocated to the link.

6. The proposed ON configuration with the selected spectrum is transferred to UE#1 by issuing a MIH_C4MS_ONN.response message.

7. To start the ON Creation (ONC), UE#1 sends a MIH_C4MS_ONC.request to the BS with the final ON configuration.

8. BS sends another MIH_C4MS_ONC.request towards UE#2 with the final ON configuration.

9. UE#2 replies with a MIH_C4MS_ONC.response message with a successful result-code indicating that the terminal is ready to establish the link.

10. BS concludes the ON creation procedure by sending a MIH_C4MS_ONC.response message to UE#1.

11. The link establishment takes place at this point.

12. Finally, the creation of the ON is notified to the infrastructure from UE#1 by sending a MIH_C4MS_ONSN.indication message.

A similar procedure is also implemented for the ON modification in case that degradation in the communication is perceived by one of the UEs. In this case, the procedure eventually triggers a new execution of the spectrum selection algorithm to modify the spectrum allocated to the link.

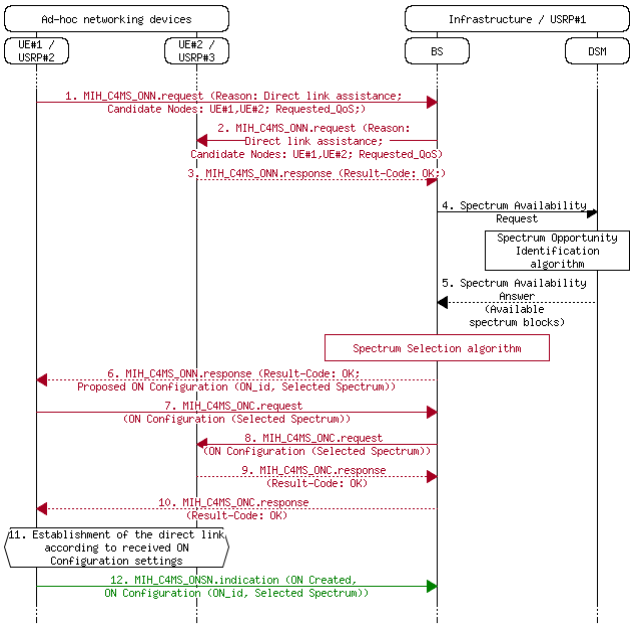


Figure 3. Implemented message exchange for the ON creation

V. VALIDATION RESULTS

In order to illustrate the testbed operation, some validation and performance results of spectrum opportunity identification and spectrum selection functionalities are presented in the following.

 ISBN: 978-1-61208-229-5

A. *Spectrum opportunity identification*

The indoor office scenario considered in this paper is illustrated in Fig. 4. The environment where the testbed operates includes the presence of two WiFi access points (AP5 and AP6) that occupy channels at 2.412 GHz and 2.432 GHz. The testbed with the ON is located in room R1.

To test the spectrum opportunity identification algorithm, the measurement procedure considered the total band from 2.4 GHz to 2.5 GHz subdivided in 1000 portions of 100 kHz. Energy detection sensing was performed for each portion during 100 ms. The threshold to detect that a portion is available is set using the following procedure: (i) the USRP antenna was replaced with a matched load (i.e., a 50 ohm resistor); (ii) the Cumulative Distribution Function (CDF) of the thermal noise was calculated; (iii) a threshold between thermal noise and signal energy was selected considering a false alarm probability equal to 1%.

Fig. 5 presents the obtained SOI for all the 1000 portions of 100 kHz averaged during a 10 minutes period. It can be observed that: (i) the spectrum portions in the ISM channels occupied by AP5 and AP6 at 2.412 GHz and 2.432 GHz, have a SOI equal to 0%; (ii) there are three groups of consecutive 100 kHz blocks with a high opportunistic index value (i.e. greater than 80%). As a result, the spectrum blocks provided by the algorithm are those indicated in Table I, considering that the maximum number of portions of a block has been set to P_{max}=290. Correspondingly, the available set of portions between 2442 to 2500 MHz with a total of 58 MHz has been split into 2 blocks.

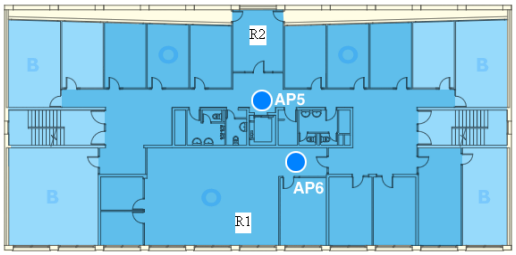


Figure 4. Considered scenario

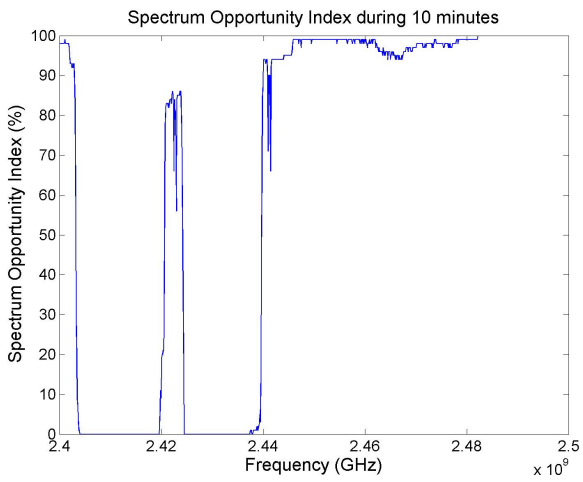


Figure 5. Spectrum Opportunity Index

TABLE I. SPECTRUM BLOCKS IDENTIFICATION

Index	Central Frequency (MHz)	Bandwidth (MHz)
1	2401.500	3
2	2422.000	4
3	2456.500	29
4	2485.500	29

B. *Spectrum Selection*

The aim of this subsection is to illustrate how the result of the spectrum opportunity identification is used to perform the spectrum selection functionality. In the scenario illustrated in Fig. 2, the terminals (i.e., USRP#2 and USRP#3) need a spectrum block to transmit data under the infrastructure (i.e., USRP#1) control. Following the procedure in Fig. 3, the allocated spectrum block is decided by the infrastructure during the ON-Negotiation procedure based on the spectrum opportunity identification executed by USRP#1 in the ISM 2.4 GHz band. The identification procedure is the same explained in the previous sub-section, but now averaging the measurements during a period of 10s and with P_{max}=200. Once the spectrum is assigned, USRP#2 is the data transmitter and USRP#3 the receiver. The experiment assumptions for the communication between terminals are given in Table II. USRP#2 periodically monitors the efficiency in the data transmission as the ratio between successfully transmitted data packets and total number of transmitted data packets including retransmissions. This is computed based on the received acknowledgements for each packet. When degradation in the communication is detected (i.e., efficiency is below the threshold of 80%), USRP#2 triggers the ON modification procedure, requesting a new spectrum block.

TABLE II. EXPERIMENT ASSUMPTIONS

Parameter	Value
Modulation	GMSK
Data Rate	256 kbps
Packet Size	1500 byte
Minimum Efficiency threshold	80%
Experiment Time	20 minutes

In the considered experiment, an additional AP has been set-up as an interference source that can be manually configured in the spectrum block allocated to the link between USRP#2 and USRP#3. Fig. 6 depicts the obtained results in one experiment. Specifically, the figure reflects the evolution of the efficiency in the communication as a function of time, in periods of 30 s. The interferer source has been activated 5 times during the experiment, leading to efficiency degradations below the threshold of 80% as can be seen in the figure. After each one of these degradations the ON modification is executed and a new spectrum block is assigned. The figure indicates the spectrum assigned to the ON link in each period of time. During the first minutes the infrastructure assigned for data transmissions the spectrum block centered at 2.422 GHz. In this period, the efficiency monitored by USRP#2 is above 80% until minute 4, when the interferer source is activated in the same frequency of the link. As a consequence, USRP#2 detects a degradation of the efficiency down to 70%. The ON modification procedure

 ISBN: 978-1-61208-229-5

leads to the assignment of the spectrum block centered at 2.452 GHz, with the corresponding increase of the efficiency. This process is repeated during 20 minutes demonstrating how the testbed is able to automatically reconfigure the assigned resources during changes in the interference conditions.

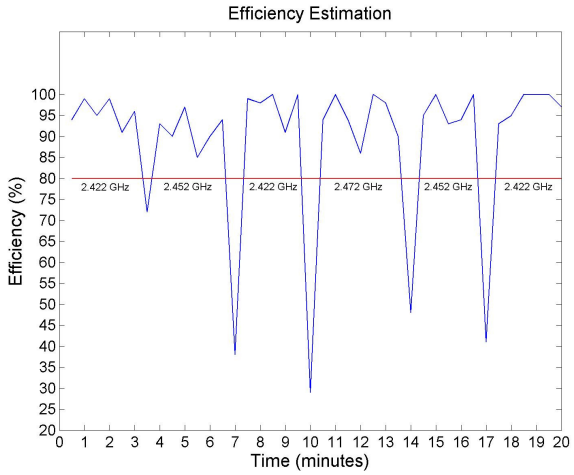


Figure 6. Spectrum Selection under changes in the interference conditions

VI. CONCLUSIONS AND NEXT STEPS

In this paper, a testbed platform has been proposed to validate the spectrum opportunity identification and spectrum selection functionalities in ON management. It is based on reconfigurable devices able to transmit and receive at different operating frequencies, and implements the necessary signaling to support ON operation. Some results have been presented to validate the implementation conducted at the laboratory and to illustrate the reconfigurability capabilities of the ON links under varying interference conditions. The developed platform constitutes a powerful tool to support the development, assessment and validation of different algorithms in real operational radio environments. Aspects related to the practicality of the algorithmic solutions can be better assessed in the testbed rather than in a simulation environment. Robustness of the proposed solutions to unpredictable radio context conditions (e.g., uncontrolled changes in the interference conditions) can also be proved in the platform. In this respect, intensive and extensive further evaluations and refinements on algorithmic solutions are expected in the near future.

ACKNOWLEDGMENT

This work is performed in the framework of the European-Union funded project OneFIT (www.ict-onefit.eu). The project is supported by the European Community's Seventh Framework Program (FP7). The views expressed in this document do not necessarily represent the views of the complete consortium. The Community is not liable for any use that may be made of the information contained herein. The work is also supported by the Spanish Research Council and FEDER funds under ARCO grant (ref. TEC2010-15198).

REFERENCES

[1] V. Stavroulaki, et al. “Opportunistic Networks: An Approach for Exploiting Cognitive Radio Networking Technologies in the Future Internet”, IEEE Vehicular Technology Magazine, Vol. 6, No. 3, pp. 52-59, September, 2011.

[2] O. Moreno (editor) “Business scenarios, technical challenges and system requirements”, Deliverable D2.1 of OneFIT ICT project, October, 2010, available at http://www.ict-onefit.eu/.

[3] L. Catalin, K. V. Rama, A. Onur, B. Dusan, and S. Ivan, “Evaluation of energy-based spectrum sensing algorithm for vehicular networks” in Proceedings of the Software Defined Radio and Dynamic Spectrum Access Technical Conference, Washington, DC, December, 2009.

[4] M. A. Sarijari, A. Marwanto, N. Fisal, S. K. S. Yusof, R. A. Rashid, and M. H. Satria, “Energy Detection Sensing based on GNU Radio and USRP: An Analysis Study”; Malaysia Int. Conf. on Comm. (MICC), Kuala Lumpur, December, 2009.

[5] R. Miller, W. Xu, P. Kamat, and W. Trappe: “Service Discovery and Device Identification in Cognitive Radio Networks”, Sustainable Development Research Network (SDRN) conference, London, December, 2007.

[6] Q. Zhao and A. Swami, “A Decision-Theoretic Framework for Opportunistic Spectrum Access”, IEEE Wireless Communications, Vol. 14, No. 4, August, 2007.

[7] H. Li, G. Zhu, Z. Liang, and Y. Chen, “A Survey on Distributed Opportunity Spectrum Access in Cognitive Network”, Wireless Communication Networking and Mobile Computing (WiCOM), Chengdu, September 2010.

[8] Q. Zhao, S. Geirhofer, L. Tong, and B.M. Sadler, “Optimal Dynamic Spectrum Access via Periodic Channel Sensing”, Wireless Communications and Networking Conference (WCNC) Hong Kong, March 2007.

[9] S.D. Jones, N. Merheb, and I.J. Wang, “An Experiment for Sensing-Based Opportunistic Spectrum Access in CSMA/CA Networks”, Dynamic Spectrum Access Network (DySPAN), Baltimore, November 2005.

[10] L. Ma, X. Han, and C.C. Shen, “Dynamic Open Spectrum Sharing MAC Protocol for Wireless Ad Hoc Networks”, Dynamic Spectrum Access Network (DySPAN), Baltimore, November 2005.

[11] ETSI, TC RRS, "Reconfigurable Radio Systems; Functional Architecture for the Management and the Control of Reconfigurable Radio Systems", TR 102 682, July 2009.

[12] J. Gebert (editor) “OneFIT functional and system architecture”, Deliverable D2.2 of OneFIT project, February, 2011, available at http://www.ict-onefit.eu/.

[13] Robin I.C. Chiang, Gerard B. Rowe, and Kevin W. Sowerby, “A Quantitative Analysis of Spectral Occupancy Measurements for Cognitive Radio”, Vehicular Technology Conference (VTC) Spring, Dublin, April, 2007.

[14] F. Bouali, O. Sallent, J. Pérez-Romero, and R. Agustí, “Exploiting Knowledge Management for Supporting Spectrum Selection in Cognitive Radio Networks,” Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM) conference, Stockholm, June, 2012.

[15] http:// www.ettus. com.

[16] http://www.gnuradio.org.

[17] ETSI TR 102 684 “Reconfigurable Radio Systems (RRS); Feasibility Study on Control Channels for Cognitive Radio Systems”, April, 2012.

[18] T. Wierzbowski (editor) “Protocols, performance assessment and consolidation on interfaces for standardization”, Deliverable D3.3 of OneFIT ICT project, June, 2012.

 ISBN: 978-1-61208-229-5

Mobile Applications for Independent Living of Isolated Elderly

Taxiarchis Tsaprounis, Katerina Touliou, Konstantinos Kalogirou, Konstantinos Agnantis, Evangelos Bekiaris
Hellenic Institute of Transport
Centre for Research and Technology Hellas
Thessaloniki, Macedonia, Greece
e-mail: Taxiarchis.Tsaprounis@certh.gr, touliouk@certh.gr, kalogir@certh.gr, kagnantis@certh.gr, abek@certh.gr

***Abstract*-REMOTE aims at an open and innovative reference architecture, based upon ontologies and semantic services that allow plug and play of existing and new services in all domains required for the independent and autonomous living of the elderly and their Quality of Life enhancement. It utilizes ICT and other key technologies in order to provide holistic services to the elderly to support their physical, social or psychological engagement and foster their emotional well being. This paper is strong evidence that telemonitoring applications may ensure maximum gain for patients as long as they are properly designed and implemented.**

Keywords-Symbian OS; Java ME; Elderly; Independent Living; Guardian Angel; Nutritional Advisor; Personal Calendar; Environmental Home Control

I. INTRODUCTION

The older population is growing at a considerably faster rate than that of the world's total population. In absolute terms, the number of elderly persons has tripled over the last 50 years and will be more than triple over the next 50-year period. The percentage of the ageing population using Information and Communication Technologies (ICT) is also rising every year [1]. Hence, future elderly users will have high ICT literacy compared to the elderly populations using similar systems to improve their overall wellbeing. The necessity of creating a platform that contains applications in all relevant domains for the benefit of the elderly user is clear and evident. REMOTE (Remote health and social care for independent living of isolated elderly with chronic condition) [2] is a Collaborative Project within the AAL (Ambient Assisted Living) joint programme which revolutionises the interoperability, quality, breadth and usability of services for all daily activities of the elderly [3].

The REMOTE approach is simple and straightforward. The core concept is direct re-usability of information across heterogeneous services and devices. The REMOTE solution is to provide foundational ontology [4] components, specifically tailored to the requirements of the applications to be covered and the services provided.

The structure of this document complies with the following description. The second section provides a general description of the project, a short state-of-the-art overview of the most crucial and relevant to the project domains and focuses on innovations over the state-of-the-art.

The third section is about REMOTE system's architecture. An overview of the functionality and the role of each module are presented.

The fourth section analyses the internal architecture of a very important component in REMOTE's architecture named Ambient Intelligence Framework.

The fifth section analyses the methodology used in the early stages of the project in order to derive REMOTE system's use cases. The sixth section presents technical specification details that a device should follow in order to be applicable for REMOTE applications' installation.

The following five sections give a detailed description of four applications which have been developed and integrated on Symbian OS platform to ensure remote access to various services related to the independent living and support of the specific target user group. Specifically, the following Java ME applications, which were developed for Symbian OS [5] mobile devices are presented:

(1) *Guardian Angel* [6] includes wearables and sensors for detecting body temperature, blood pressure, heart rate, human posture and motion/acceleration recognition and sends alerts to the healthcare professional in case an abnormal measurement is detected.

(2) *Nutritional Advisor* [6] offering everyday tips on nutrition, weekly menus and recipes according to the user's needs and preferences.

(3) *Personal Calendar* [6] for scheduling/managing daily tasks for the elderly (concerning nutrition, medication and to-do lists) under the unobtrusive supervision of their carers.

(4) *Environmental Home Control* [6] for interacting with home appliances and monitoring the house's status remotely.

All the above applications are interoperating with each other by the use of the user profile mechanism and store patient's monitored data to remote servers which can only be accessible by the professional healthcarers with the use of web applications.

The pre-last section presents the evaluation phase results at determining the usability and acceptance of four of REMOTE's applications and the last section is about conclusions that were gathered during the evaluation phase of the project.

II. REMOTE CONCEPT, STATE-OF-THE-ART AND INNOVATIONS

REMOTE EU project can be divided to REMOTE ontology, REMOTE platform and REMOTE applications.

REMOTE ontology basis is a set of existing ontologies related to each application, which was initially planned to be developed for the needs of the project, i.e., health monitoring, physical activity, mental exercise, nutrition,

 ISBN: 978-1-61208-229-5

communication and calendar tasks. After its finalization the ontology was evaluated by the use of the "Competency Questions" process [7]. REMOTE platform is a framework that allows integration of single services using ontological layering. Services are integrated into the system through a "service-ontology alignment" process. Specifically, the service alignment is realised by the service provider through an alignment and anchoring tool, which is provided through REMOTE and semantically matches (aligns) ontological concepts to web services structural components, i.e., their I/O (input/output) parameters. REMOTE applications, which are invoking the integrated services, are provided to the elderly through the main menu.

Because of the fact that any service following certain specifications may be integrated in the system, REMOTE is considered to be an open reference architecture that allows data and (external) services fusion.

Primary users of REMOTE are the elderly with specific user conditions, especially those living in rural and isolated areas. Secondary users are the professional health carers with tools for continuous monitoring run-time and history patient data. Tertiary end-users are the service providers who are integrating their services into REMOTE platform with the use of the alignment tool.

Taking into account all previous research projects REMOTE's applications have used and further developed results from, i.e., for elderly people: ASK-IT [8] and MAPPED [9], for mobility issues: COGKNOW [10] and EMPOWER [11], to foster daily activities: SOPRANO [12], INHOME [13], OLDES [14] and AMIGO [15]. The most significant innovation in REMOTE is the multi sensor approach, i.e., body and home environmental sensors, performing a data fusion of their various inputs, combined with expert knowledge and individual user information. Particularly, new mechanisms have been defined to hide the complexity of the various sensor network environments and user interface adaptation algorithms have been created that take into account the users' needs and preferences in order to adapt the user interface according to situational and technical context of interaction. The use of ontologies in REMOTE is an innovative feature that gives added value to the platform. The ontologies assist in sharing common understanding of the structure of information among people or software, they enable reuse of domain knowledge and make it easier to analyze it.

A fundamental aspect of REMOTE system (sensors, devices, software, etc.) is its scalability, flexibility and adaptability "characters" that help to be easily integrated into existing set-ups and contexts. REMOTE prototypes and technology-based solutions are well-adapted to the respective diagnosis, prevention and treatment opportunities that can be attained while allowing the elderly to stay "at home" (detecting signs, symptoms, and risk factors; monitoring cure processes; etc.). The REMOTE system developed new elderly-oriented human-machine interaction paradigms, new systems for monitoring users at home, e.g., Dehydration level measurement and/or when they are on the move, e.g., Guardian Angel application. Moreover, the system detects important events (health risks, daily activity and behaviour).

III. REMOTE'S ARCHITECTURE OVERVIEW

The most important components of the REMOTE architecture are briefly described in the following list (Fig.1).

A. Ontology Repository (OR)

It is the technological layer that supports the Ontologies storage and management [16].

B. Common Ontological Framework (COF)

The COF defines a formal specification of ontology modules, and how they relate. The COF defines a methodology and best practice for ontology construction. It makes possible to define an ontology and facilitate and optimize the integration of new emerging ontologies.

C. Content Anchoring and Alignment Tool (CAAT)

This tool aligns the functionality of the provided web service through its Web Service Description Language (WSDL) [17] file with the ontologies stored in the OR. The concepts of the same or different application areas, after being aligned with the appropriate ontological concepts are ready to be used seamlessly through the CCM. The purpose of the Concept Anchoring and Alignment tool is to allow service providers insert their web services into the REMOTE framework.

D. Content Connector Module (CCM)

CCM receives a request for service by the end-user (client) application via the Ambient Intelligence Framework (AmI) and invokes the appropriate service that returns the requested content to the client [18].

E. Ambient Intelligence Framework (AmI)

The role of AmI framework [18] in REMOTE system is to provide seamless interactivity between REMOTE applications and the Content Connector Module. The Content Connector Module exposes its functionality as a web service and it is invoked by the AmI through a web service client.

F. User Profile

It contains all the context information related to a specific user. If a REMOTE component needs to retrieve some information related to the user context but out of its own scope, it should make a query to this user profile [18].

 ISBN: 978-1-61208-229-5

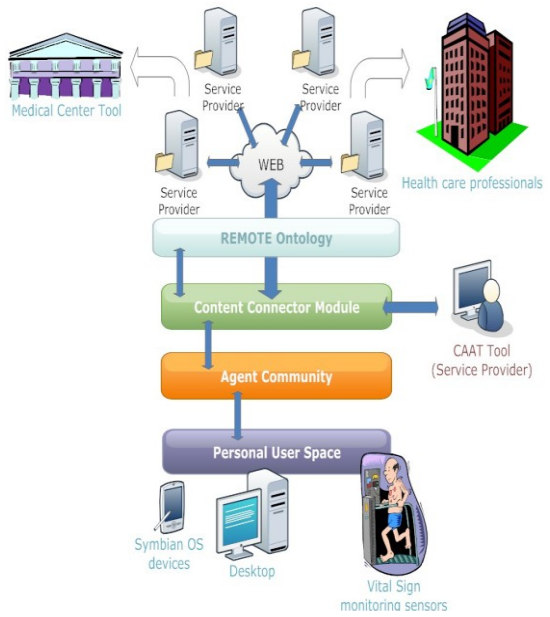


Figure 1. REMOTE Architecture

IV. AMBIENT INTELLIGENCE FRAMEWORK OVERVIEW

The AmI framework consists of four agents [19]. It directly communicates with the Content Connector Module through the Service Provider Agent and with the device through the Dialog Manager Agent. The User Profile Agent is responsible for accessing the User Profile Repository. The AmI internal communication and the user profile repository location are depicted in Fig.2.

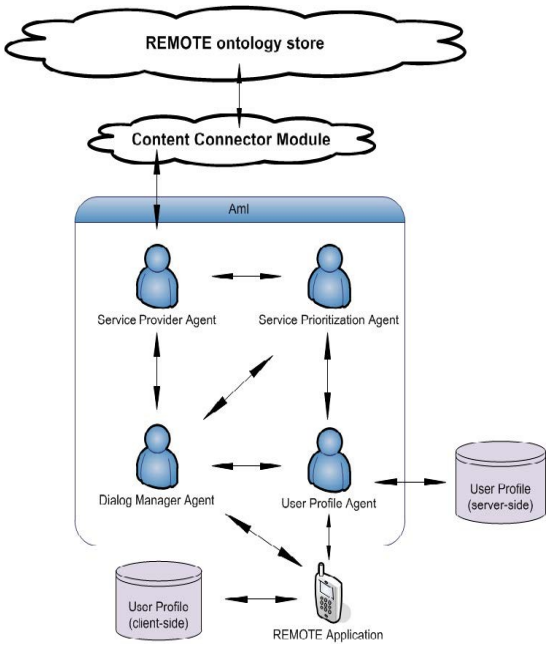


Figure 2. Ambient Intelligence Framework

The **Dialog Manager Agent** role is to contribute to a distinct and clearer AmI internal architecture. It resides on the client side and handles the communication between the mobile and the REMOTE server part.

The Content Connector Module of the REMOTE framework performs semantic search among suitable available web services, in order to satisfy the user request. The results produced as outcome of the search mechanism are then fed to the **Service Prioritisation Agent**, whose role is to provide a ranking of the returned services according to the specific needs, preferences and habits of the user. It actually implements a low-level information filtering process, thus providing the most valuable services to the end user. It prioritizes web services by taking into account meta-data received from CCM. Concerning the meta-data parameters currently used for service prioritization it is important to define that they are distinguished between parameters used for filtering services and parameters used after the filtering process for prioritizing them. In case the meta-data values of a service do not match the filtering parameters the corresponding web service is not taken into consideration in the prioritization process.

The filtering meta-data parameters are the following:

- *Language* (Comparison with the user's preferred languages)
- *Country* (Comparison with the user's country)
- *City* (Comparison with the user's city)

The prioritizing meta-data parameters are stated below:

- *Age Category* (young elderly, elderly, old elderly)
- *Max Accepted Cost*
- *Impaired Category*

The **Service Provider Agent** is responsible for the AmI-CCM communication.

The **User Profile Agent** is the agent who has direct access to the User Profile Repository in order to obtain data that will be used for filtering the returned services and optimising the result returned to the user. It is responsible for the storage and retrieval of all these profile properties of the user that are needed by REMOTE applications.

V. METHODOLOGY

Use Cases were created early in the project and the process for finalising them followed several adjustment phases. Firstly, a State-of-the-Art search for identifying relevant systems and services in the domain of Telemedicine applications was carried out. Secondly, an online survey-via REMOTE website - was conducted in eight European countries and a set of interviews has been carried out in five of them for the extraction of user needs and requirements. Based on these findings, the initial Use Cases were developed. Both face-to-face interviews and online surveys were conducted in order to gather as much information as possible about the needs of end users.

A total of 266 individuals from 6 countries (Spain, Israel, Greece, Germany, Norway and Italy) were surveyed via face-to-face interviews, as well as through online questionnaires. As a result of the whole process, 41 use cases have been defined. These use cases were the cornerstone of the architectural design and development of the REMOTE

 ISBN: 978-1-61208-229-5

system, which has been continuously improving by taking into consideration each one's of the evaluation phases' results.

VI. DEVICE TECHNICAL SPECIFICATIONS

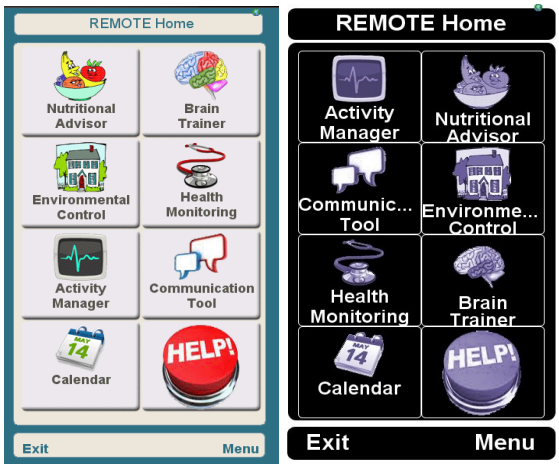
In terms of application development for Symbian OS devices, Java ME (Java Mobile Edition) technology [20] has been used. The Lightweight User Interface Toolkit (LWUIT) [21], which is a versatile and compact API for creating attractive application user interfaces for mobile devices, has been applied for the implementation of the REMOTE mobile user interface.

REMOTE application can be downloaded to CLDC (Connected Limited Device Configuration) 1.1 [22] and MIDP (Mobile Information Device Profile) 2.x [23] devices.

The devices that may be used for REMOTE installation should support JSR (Java Specification Request)-75 (specification that standardizes access in Java on embedded devices -such as mobile phones and PDAs- to data that resides natively on mobile devices), JSR-179 (GPS-Global Positioning System- functionality), JSR-135 (extends the functionality of the JME platform by providing audio, video, and other time-based multimedia support) and JSR-172 (enables Java ME devices to be web service clients). The use of a Wi-Fi enabled mobile device is recommended because of the faster responses the user receives when invoking a service through a WLAN (Wireless Local Area Network) rather than 3G/GPRS/UMTS connection and the fact Wi-Fi can often fill-in some of those dead spots or signal losses (i.e., inside the house).

VII. MAIN MENU OVERVIEW

When launching the REMOTE application, the user is prompted to insert his/her username on a login screen. The initialisation of user profile is taking place in the Medical Contact Centre (administrative web application) and its medical details are filled with face-to-face communication with the health care professionals. Main Menu (Fig. 3) displays only the registered applications of the logged-in user, which are also defined during the user profile initialization process (Medical Contact Centre). Personal preferences acquired by the user profile are taken into account for the automatic selection of the appropriate theme, font size and language. REMOTE supports eight different languages (i.e., English, Greek, Romanian, Spanish, German, Norwegian, Italian and Hebrew), four different themes for different user chronic conditions and three font sizes.



(a) (b)

Figure 3. REMOTE Main Menu: (a) blue theme, (b) high contrast theme

VIII. GUARDIAN ANGEL

Guardian Angel application aims to support the users on the move via monitoring technologies based on state-of-the-art wearable and mobile systems, throughout their daily activities. For this purpose, sensor-enhanced devices are incorporated for the unobtrusive monitoring of various vital parameters such as heart rate, breathing rate, posture and activity, skin temperature, blood pressure, weight, etc. According to the definition of personalized monitoring schemas, appropriate alerts can be generated so as to assist the patients in avoiding or overcoming hazardous health conditions or situations, i.e., arrhythmia, high blood pressure and fall. The system is primarily targeted at patients with chronic diseases such as hypertension and Parkinson's disease.

The Guardian Angel Mobile Application is mainly consisted of five distinct services, deployed on the user's mobile device: a) The sensors communication, b) the sensor data handling, c) the emergency management, d) the vital signs management, and e) the proximity.

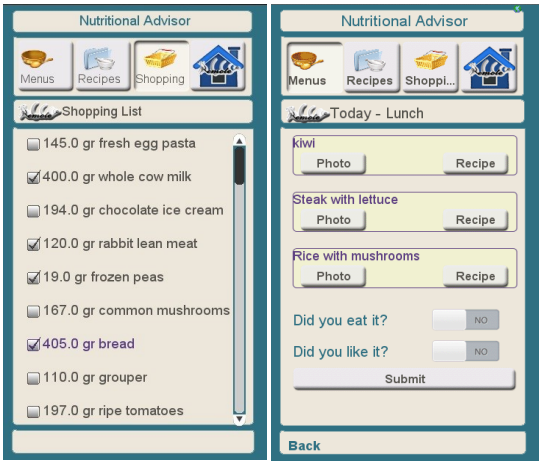
The sensor-enhanced devices can communicate via Bluetooth with the patient's mobile device, forming in this way a body-sensor network around the user for health monitoring purposes. Guardian Angel mobile application utilizes the multi-sensing wearable strap Zephyr BioHarness [24]. BioHarness is a device particularly suited for monitoring the heart rate, breathing rate, activity, posture and skin temperature.

The patient is provided, by the initial application screen, with instructions for wearing the strap properly, since this is a critical requirement for the efficient monitoring of his/her health status. Clicking on "What is measured" button, the patient can view his/her measurement values as well as their corresponding thresholds. If a measurement is out of range then proper alerts are generated.

 ISBN: 978-1-61208-229-5

(a) (b)

Figure 4. REMOTE Guardian Angel: (a) Vital sign measurement, (b) Emergency alert



(a) (b)

Figure 5. REMOTE Nutritional Advisor application: (a) Shopping List, (b) Lunch menu display – Nutrition telemonitoring

After pressing the "Start" button, communication with the BioHarness sensor is initialized and the patient's parameters - heart rate, breathing rate, skin temperature and posture (i.e., controlling for fall detection) start being monitored (Fig. 4a).

In case of alert (Fig. 4b), the user is notified and an orange alert is transmitted to the emergency service of the Medical Contact Centre. If the alert persists for a period of 55 seconds, a red alert is triggered and sent to the Medical Centre so that appropriate actions are routed. The patient will be notified when the measurement value is back to safe levels. The patient's monitoring values (received by the BioHarness device every second) are sent to the back-end system periodically and health professionals are informed of his/her current health status. The patient is notified in case of loss of internet or Bluetooth connection.

IX. NUTRITIONAL ADVISOR

The main purpose of nutritional profiling is to detect nutritional risks and preferences for users and the nutritionist can determine the best nutritional plan taking into account their health condition. Nutritional Advisor application aims at controlling and influencing the nutritional habits of the elderly. It provides appropriate data that facilitate their nutritional-related daily habits (i.e., grocery shopping and cooking activities). In particular, the shopping list feature assists the elderly in purchasing everything he/she needs for preparing the meals of the day/week and the recipes feature provides everything that the user may need in order to prepare a meal such as meal photos, detailed instructions and precise quantities of each ingredient needed (Fig. 5a).

The system recommends to the user a daily menu (Breakfast, Lunch and Dinner) based on the nutritional plan prepared by the nutritionist, which takes into account the calories of each menu, the user's medical profile and health status (allergies, medications and diseases) and also the user's activity level (Fig. 5b).

All the nutritional content provided by the Nutritional Advisor service, is customized for each specific user, taking into account their user profile, which gathers the likes, needs and requirements of the elderly user. Due to the fact that nutritional content is totally depended on the service provider, no extra details can be given about it. The exchange of information between the Nutritional Advisor and other REMOTE applications like the Guardian Angel or the Activity Coach improves both the nutritional service, as well as the other applications connected to it.

X. PERSONAL CALENDAR

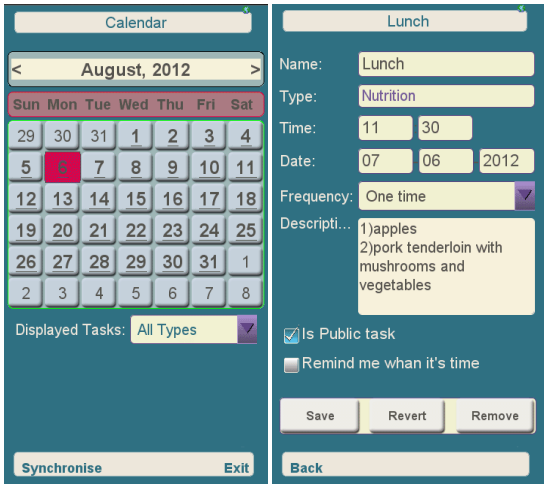
The personal calendar application for REMOTE is used for scheduling and managing the daily tasks of the elderly under the unobtrusive supervision of the caregiver. The application enables the user to keep a good schedule of their life and activities and also serves as a memory aid to assist their memory.

The application offers four different kinds of functionalities: (a) calendar management, (b) task management, (c) notifications and (d) integration with other applications.

Calendar management has to do with the daily schedule of the user. A standard calendar view is provided and the user can check and navigate through days and months (Fig. 6a). The tasks that have been added to the calendar are visible and can be examined in detail. These can be either user-created tasks or notes by the healthcare professional or the caregiver. It contains all entries that are relevant to the current day (e.g., including any suggested activities), medication intake and calendar tasks and notes (Fig. 6b).

Task management has to do with personal or third-party tasks that concern the user. The application can show the list of tasks that have been added to the calendar and offers services for adding or deleting them.

 ISBN: 978-1-61208-229-5



(a) (b)

Figure 6. REMOTE Personal Calendar application: (a) Calendar panel, (b) Lunch stored as nutrition task

(a) (b)

Figure 7. REMOTE Home Environmental Control application: (a) Available rooms, (b) Integrated sensors/devices

These services are also offered to other applications that may need to manage the user's tasks, such as an application for the professional, allowing him to add medicine-related tasks for the user or the nutritionist's application that displays the daily meals of the elderly as calendar tasks.

XI. HOME ENVIRONMENTAL CONTROL

The Environmental Control application enables the users to control household appliances from a remote distance. The user can either be at home controlling devices of a room from another room or from a place outside his house. This interaction is achieved through actuator devices such as switch or dimmer switches. Home monitoring is also possible through the use of sensors which have been installed in different places inside the house. Temperature, humidity, luminance are some of the various measurements that can be monitored in order to make the user feel more secure about the environment s/he lives in. A home automation lab was built at CERTH's (Centre for Research and Technology Hellas) premises in order to demonstrate and validate the functions that are supported by the REMOTE system. The home environment is very heterogeneous composed of different application areas, devices communication standards, user needs and wants. Moreover, many incompatible systems are available on the Market. This leads to many incompatible "communication islands" at home, with no interoperability and no "overall interaction". Therefore, the Home Environmental server should follow a "multi-standard-approach", facilitating to control devices using different communication protocols on various busses at the same time. This defined some technical requirements to the architecture of the server:

- It needs to "know" the busses, its protocols and how to use them.
- Each device has to provide a communication protocol, able to receive commands and/or answer requests.
- The server needs to "know" the devices that should be controlled.

Each device is described by a Device Profile using XML, defining the communication protocols and their parameters and containing the names of the devices and other device properties [25].

All devices are using wireless communication protocols based on 868MHz ISM band, i.e., FS20, HMS, S300.

Clicking on the "Environmental Control" button on the REMOTE main menu displays the Environmental Control menu which consists of a horizontal menu with the "Rooms" and "Home" buttons and a vertical menu where each button represents a room (Fig. 7a) except the last which updates the sensor values and the pre-last one which displays the energy that is currently consumed by the devices which are connected to the system.

Clicking on any of the room buttons (Living room, Kitchen or Bedroom) displays the devices connected to the room and the current sensor values. The living room electrical appliances can be seen on Fig.7b. On the left side of the central panel the user can check the room's temperature and humidity, whether there is any motion inside the house, if the door is opened and if the room is dark or light.

XII. RESULTS

A user-centred approach was adopted early in the REMOTE project aiming to accommodate the needs of patients with chronic diseases that might live in isolated regions. An iterative testing cycle was conducted with experts in the field of usability testing and developing. Overall, applications were regarded as sufficient and accepted. Comments for both functionalities and Graphical User Interface (GUI) were uploaded to the Mantis tool [26] for responsible partners (i.e., developers) to have access. It was necessary- as first rectification step to match experts'

 ISBN: 978-1-61208-229-5

prioritisation to developers' prioritisation for pragmatic improvements. Respective bugs, errors and issues were dealt with prior the final evaluation phase with real users. The evaluation phase aimed at determining the usability and acceptance of the mobile applications.

A. Participants

17 female (57.8±6.61 years old) and 13 male (59.8±4.12 years old) users participated in the pilot study. All users were derived by an existing user database. 27 users suffered from chronic diseases (e.g., hypertension, diabetes, arthritis) and 3 were healthy elderly. Most users live with their spouses (25/30). All users are adequately familiar with mobile phones. Participants provided written consent prior participation and received compensation.

B. Main Findings

Evaluation was based on tasks' completion derived by two fictional scenarios. Each session lasted approximately two hours with short breaks and was audio recorded in order to gain as much insight as possible from "think aloud" processes. Users had constant support by two facilitators throughout testing.

Post-task analysis showed that overall success to completed appointed tasks to users was high (Fig. 8) for all mobile applications tested. Task analysis was based on steps and time taken to successfully complete the appointed task. At least three tasks were completed by each user per tested mobile application.

Calculation of task completion success rates was based on both users' and facilitators' ratings. Users had to state if they thought they successfully completed each task and at the same time the facilitators recorded their own rating (i.e., success/partial success/failure). The average score from both facilitators and user was the overall user success rate. Increased success in completing the tasks was found for the health monitoring application. However, fewer steps were required to complete the respective tasks for this application and it might have affected the overall success rate for this specific mobile application. Nevertheless, success rates are all above 90% and are quite impressive taking into consideration that most users (83%) were not acquainted with touch screen mobile phones. These findings support both the easiness and learnability of the mobile applications developed within the framework of REMOTE but also their increased potential for deployment and penetration to existing telemedicine and health mobile applications with minimal instructions and training.

User acceptance ratings ranged from -2 (negative) to 2 (positive). The two extremes were defined by the content of each questionnaire item (e.g., unpleasant-pleasant). Mean user acceptance ratings ranged from 1.62 (SD: ± 0.32) to 1.76 (SD: ± 0.26) (mean and standard deviation for Nutritional Advisor and Environmental Control, respectively). Mean user acceptance scores for Health Monitoring and Calendar were similar (1.69 ± 0.41 and 1.72±0.22, respectively). Mean acceptance scores were high for all four mobile applications.

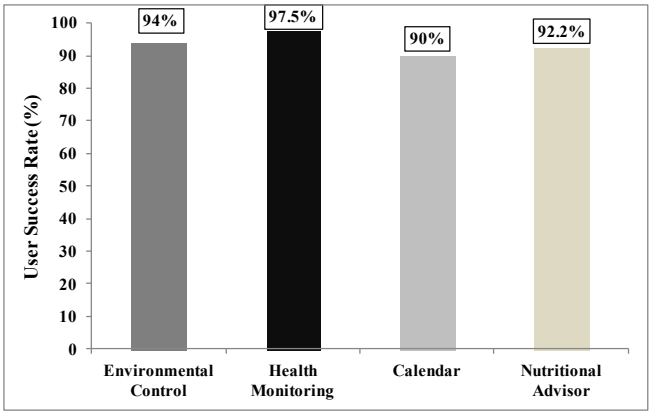


Figure 8. User success rates (%) per tested mobile application

Fig. 9 depicts Usability scores [27] for each application tested. Overall, all mobile applications were perceived as useful with higher percentages found for both Environmental Control (78.5%) and Calendar (77.5%) applications. Lower ratings were recorded for Health Monitoring (67.3%) and Nutritional Advisor (66.6%) applications probably because users were not familiar with these types of applications at all. Hence, the differences might lie in the purpose of usage and content of these applications. In addition, for Environmental Control the results were evident (e.g., they switched on a light and they could see it). On the other hand, they received information about a vital sign (e.g., heart rate) but they did not know if this was true or not. In other words, the trust to the system was higher for some applications when compared to others. Moreover, the content for Health Monitoring and Nutritional Advisor was increased compared to the other two applications for both complexity and appearance (i.e., Information appearing on the screen).

Overall usability (72.5%) is adequate and above average but it also shows that further improvements could be made resulting into an even more usable system. Therefore, decreasing complexity and amount of information displayed at screen might increase usability.

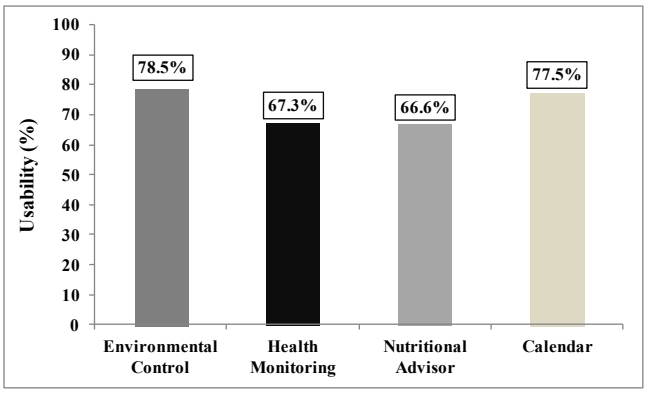


Figure 9. Usability scores (%) per tested mobile application

XIII. CONCLUSIONS

The REMOTE mobile system provides a flexible solution for elder patients for services ranging from health

monitoring to environmental control. Technical advancements are required for these services to be available and accessible to the older populations.

The REMOTE system has been developed within a research framework and it is not yet a marketable product. Therefore, its adaptation to the commercial needs should be implemented with the assistance of relevant stakeholders (i.e., telecommunication service providers and mobile device vendors) in order to be optimised when a marketable version will be available.

Overall, all four mobile applications were regarded as usable, easy to learn and desired by most users. The latter is highly dependable to the affordability of this system as it is an important prerequisite for most elder users.

Finally, it is worth to be indicated that research on holistic approaches to providing services to isolated elderly and patients via New Technologies is nowadays an essential research tool that could result into a valuable assistive and supporting product for better quality of life [28].

ACKNOWLEDGMENT

Work presented in this paper was achieved in REMOTE research project, which is cofounded by the European Commission, under the Ambient Assisted Living (AAL) joint program (AAL-2008-1-147).

REFERENCES

[1] Pazienza M.T., Stellato A., Vindigni M., and Zanzotto F.M., "XeOML: An XML-based extensible Ontology Mapping Language", Workshop on Meaning Coordination and Negotiation, held in conjunction with 3rd International Semantic Web Conference (ISWC-2004) Hiroshima, Japan, November 8, pp. 83-95, 2004.

[2] *REMOTE EU project*, http://www.remote-project.eu/ [retrieved: October, 2012]

[3] Rocker C., Ziefle M., and Holzinger A., "Social Inclusion in AAL environments: home automation and convenience services for elderly users", Proceedings of the international Conference on Artificial Intelligence (ICAI'11), vol.1, Las Vegas, NV USA, pp. 55-59, July 18-20, 2011.

[4] Uschold M. and King M. "Ontologies: Principles, Methods, and Applications. Knowledge Eng. Rev., vol. 11, no. 2, pp. 93-155, 1996.

[5] *Symbian OS,* http://licensing.symbian.org/ [retrieved: October, 2012]

[6] Tsaprounis T., REMOTE D4.2, "Mobile and portable self-care services and applications", 2011.

[7] Gómez-Pérez A., Fernández-López M., and Corcho O., "Ontological Engineering", Springer-Verlag London Limited, London, UK, pp. 119-120, 2004.

[8] *ASK-IT EU project*, http://www.ask-it.org/ [retrieved: October, 2012]

[9] *MAPPED EU project*, http://services.txt.it/MAPPED/ [retrieved: October, 2012]

[10] COGKNOW EU project, http://www.cogknow.eu/ [retrieved: October, 2012]

[11] *EMPOWER EU project, http://www.ep-empower.eu/* [retrieved: October, 2012]

[12] *SOPRANO EU project*, http://www.soprano-ip.org/ [retrieved: October, 2012]

[13] INHOME EU project, http://www.ist-world.org/ProjectDetails.aspx?ProjectId=fdb62df32f954628a8308a0de08cbf6f&SourceDatabaseId=7cff9226e582440894200b751bab883f [retrieved: October, 2012]

[14] OLDES EU project, http://www.oldes.eu/ [retrieved: October, 2012]

[15] *AMIGO*, http://www.hitech-projects.com/euprojects/amigo/ [retrieved: October, 2012]

[16] *REMOTE Ontology Repository*, http://orate.iti.gr/ [retrieved: August, 2012]

[17] *WSDL*, http://www.w3.org/TR/2007/REC-wsdl20-adjuncts-20070626/#_http_binding_default_rule_method [retrieved: October, 2012]

[18] Tsaprounis T. and Giannoutakis K., REMOTE D2.2, "REMOTE AmI framework and agents", 2010.

[19] Bellifemine F., "Developing multi-agent systems with JADE", , Wiley, Liverpool University, UK, 2007

[20] *JavaME*, http://www.oracle.com/technetwork/java/javame/index.html [retrieved: October, 2012]

[21] *LWUIT*, http://lwuit.java.net/ [retrieved: October, 2012]

[22] *CLDC*, http://java.sun.com/products/cldc/overview.html [retrieved: October, 2012]

[23] *MIDP*, http://www.midp.net/ [retrieved: October, 2012]

[24] *Zephyr*, http://www.zephyr-technology.com/bioharness-bt [retrieved: October, 2012]

[25] Kalogirou K. and Telkamp G., "An ontological framework for the elderly to control their home environment", Durban, South Africa, 19-21 May, IST 2010.

[26] *Mantis Tool*, http://www.mantisbt.org/ [retrieved: October, 2012]

[27] Brooke J., SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland. Usability Evaluation in Industry. London: Taylor and Francis, pp. 189--194 , 2006.

[28] Pare G., Jaana M., and Sicotte C., Systematic Review of HomeTelemonitoring for Chronic Diseases: The Evidence Base: J Am Med Inform Assoc., vol.14, pp. 269-277, 2007.

 ISBN: 978-1-61208-229-5

User Experience Evaluation in the Creation and Use of Graphical Passwords for Authentication in Mobile Devices

Claudia de Andrade Tambascia, Ewerton Martins Menezes, Alexandre Melo Braga, Flávia de Melo Negrão
CPqD Foundation
Campinas – SP, Brazil
{claudiat, emenezes, ambraga, fnegrao}@cpqd.com.br

***Abstract* — This article aims at present the results of a user experience evaluation for the creation and use of graphical passwords on mobile devices as a way to improve usability and security aspects in authentication. The authentication method proposed was defined in a project called Multimodal Biometric and Graphical Authentication for Mobile Devices. These assessments were carried out with thirty users during a period of fifteen days using a prototype that offered a repertory of eighty icons divided into four categories. All users were able to remember their eight-icon password and claimed having a good use experience with this authentication method.**

Keywords - User experience; graphical passwords; mobile authentication.

I. INTRODUCTION

One of the main goals of the use of graphical passwords in authentication processes is to increase the usability of the interaction, facilitating the memorization of passwords for users, ensuring a greater retention of information. This feature relies on the results of various studies and experiments [1][2][3][4] that show the human brain finds it easier to recognize and remember visual information comparing to textual information.

As a recognition-based technique, the graphical authentication with icons demands less cognitive load than recall techniques and tends to increase the usability, the security and the user performance, besides being especially appealing in the mobile context, where pointing to a region of the screen tends to be much easier than typing text.

While some recognition-based systems use faces [5], assuming that the brain has got a special ability to recognize them, other systems use abstract images [6], which are stronger from a security point of view, due to their difficulty of describing. Nevertheless, the use of icons brings a better compromise between usability and security, once it facilitates mnemonic strategies and, consequently, memorization.

The security level offered by such systems depends on many factors, such as the size of the repository available to the user, the password length, the input method, and the icons themselves, which must, ideally, present similar probabilities of choice avoiding possible attacks.

This paper will describe the results obtained in the evaluation of the quality of user experience in the creation and use of graphical passwords for authentication on mobile devices. For this evaluation, a prototype was developed to allow the experience in a real context of use, considering aspects of usability, intelligibility and memorization strategies. This prototype is part of Multimodal Biometric and Graphical Authentication for Mobile Devices (BIOMODAL) project whose main goal is to develop functional prototypes of biometric multimodal authentication and graphical authentication for mobile communication devices.

Section two will present a description of the prototype that was developed for this evaluation, followed by section three that will present the methodology used for the user experience evaluation. Section four will present the main results observed during the exploration, the creation and the strategies of memorization applied followed by section five with the analysis of use data for the different groups of users and interviews. The section six will present the main conclusions and findings related to graphical authentication.

II. RELATED WORKS

As a knowledge-based authentication technique, the graphical authentication requires the user to enter a shared secret as an evidence of their identity. This authentication scheme has been proposed as an alternative to text-based password for over one decade [7]. Surveys on the field [8][9] review some graphical password systems from the usability and security perspective, but Robert Biddle's survey [7] provides us with a comprehensive review of the first twelve years of published research on graphical password systems. According to this study, graphical passwords scheme can be classified in three main categories: based on recall, recognition and cued-recall.

Graphical passwords with icons are a type of recognition-based systems where users are asked to memorize a repertory of images during password creation, and then recognize their images from among decoys to authenticate. Proposed recognition-based systems use various types of images, most notably: faces, random art, everyday objects, and icons. Renaud [10] discusses some specific security and usability considerations, and offers usability design guidelines for recognition-based systems.

In the literature it's possible to find studies such as Antonella [11] that compares the a PIN password against three different graphical passwords schemes in two user studies with 60 participants and verified that graphical

 ISBN: 978-1-61208-229-5

password can be a solution to some problems related to knowledge-based authentication, but poor design can eliminate pictures superiority effect in memory.

Darren study [12] evaluated some security and usability aspects involved in graphical passwords with icons using faces and showed the importance of some rules in the password selection in graphical schemes due to the highly correlation between race and gender of the user.

And due to the specially appealing of graphical password in the mobile context, where typewritten input is less common than pointing at the screen [13] we found studies like the Dunphy's one [14] that evaluate two versions of a recognition-based system on mobile devices where the images used in the application were provided by the users himself.

Although there are many academicals studies about graphical passwords with icons only a few commercial products are available in the market such as Passfaces and the over take-up is low [14] [15].

III. PROTOTYPE FOR CREATION OF GRAPHICAL PASSWORDS

The evaluation of user experience with graphical passwords was performed by the means of a prototype developed and installed in a Samsung Galaxy S smartphone, with 4 inches (10 centimeters), resolutions of 480×800 pixels and Android operation system version 2.2, as shown in Figure 1.

Figure 1 - Prototype installed in a smartphone

The prototype enabled the evaluation to be conducted in real devices, allowing evaluating not only graphical authentication parameters but also identifying interaction requirements inherent to mobile platform. Users were able to create and use graphical passwords in the device that automatically registered: time of use, number of page scrolls, use of "clear" function and number of authentication attempts, which are parameters needed to measure the quality of user experience.

For the creation of the graphical passwords, a repertory was produced based on the results presented on [16] that adopted mnemonic strategies to favor the use of episodic memory in the moment of password creation. The repertory was composed of eighty icons, displayed in 4 screens of twenty icons each, classified in categories: people, objects, means of transportation and context/place, as shown in Figure 2.

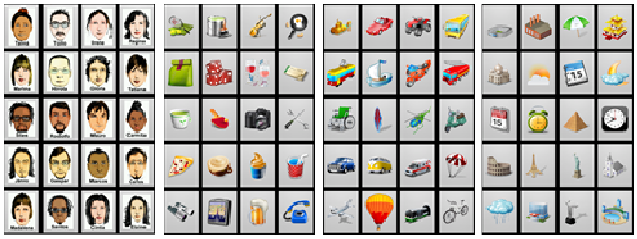


Figure 2 - First repertory of icons

For the usability study it was designed a simple prototype, which not bound by strong password policies. This design decision allowed the study of users unsafe behaviors. A commercial authentication solution based on this technology not only must implement stronger passwords (longer), but must provide security policies that inhibit the passwords considered that would facilitate easy and dictionary attacks. Only for an example, a graphical password cluttered with 9 icons in a grid of 80 icons is equivalent of a password of 8 letters, and a password of nine icons arranged in a grid of 57 icons is equivalent of 8 size alphanumeric password with uppercase letters, lowercase letters, numbers and special characters.

As well as in the use of alphanumeric passwords, it was possible to observe that some of users tried to eliminate apparent or implied variables of the repertories to facilitate their process of memorization. Thus, to enhance the quality of the passwords, besides not allowing choosing repeated icons, additional restrictions were imposed in the moment of creation: not choosing more than three icons in each screen to avoid that all icons are in the same page or belong to the same category.

From the application interface point of view, this evaluation sought to identify and validate the adequacy of the quantity of icons presented per screen, easy and intuitive navigation, and clarity of the actions and feedbacks offered to the users.

IV. PROTOTYPE EVALUATION METHODOLOGY

In order to deeply explore the process of graphical password creation, the user sample was divided into three groups with distinct methods of creation. The first group was able to create the password according to their own personal preferences; the second group was suggested a random theme (adventure, romance, work, celebration, contretemps, tourism, luck and trouble) and the third group received a password that was randomly generated by the application with two icons in each screen.

Training is a factor of great influence in the memorization and performance of any password. In order to understand of the impact of this variable in graphical passwords, half of the participants went through a training session where they used the password three consecutive times right after creating it.

A total of thirty users representing general population participated in the test divided into six groups taking into

 ISBN: 978-1-61208-229-5

account the interaction with or without password training, user created password with or without a suggested theme and the memorization of randomly generated passwords by the application and given to the users right before the first interaction, as shown in Figure 3.

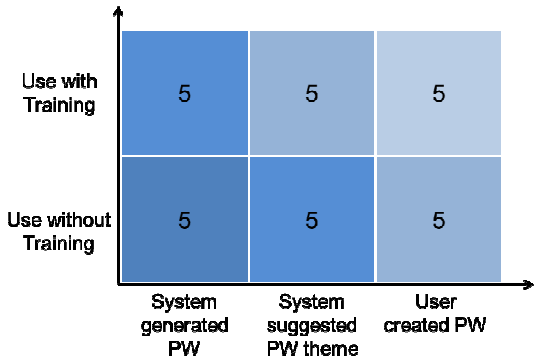


Figure 3 - User distribution according to established groups

The evaluation consisted of six stages, where some stages were common to all groups and others, however, were specific to certain groups. The stages considered were:

- Exploration: in this stage users were requested to visually explore the repertory of icons displayed in the eye tracking device. No interaction with the mouse or keyboard was required since the screens were presented for a preset time.
- Creation: the creation of the password of eight icons was performed in the mobile device instantly after the repertory exploration. A total of twenty users underwent the process of password creation, and ten of them were suggested a theme to help create the password.
- Password generation: the password was generated in an application developed to randomly choose two icons from the repertory in each screen, composing an eight icon password presented to the user right after exploration.
- Training: consisted of entering the password three consecutive times in the mobile device. The user did not need to enter the correct password every time, but after each interaction the system informed if the password was correct or not.
- Password effective use: the graphical password was used four times during the evaluation, in increasing time intervals from the creation until the last use.
- Interview: it was conducted at the end of the evaluation, where some questions were asked to the participants to verify their perception regarding the interaction with graphical passwords. The following aspects were treated: experience, difficulty of use, positive and negative issues, tendency to replace alphanumeric passwords for graphical, passwords created and memorization strategies, category harder to recall, category most liked, experience with smartphones and the mobile device model owned by the participant user.

To evaluate the performance between different groups of users, quantitative and qualitative parameters were collected. The quantitative parameters considered the time of creation and uses of the passwords, the success and failure rates, the number of attempts in each authentication, the use of “clear” function, navigation between screens, icons chosen and icons looked at. The qualitative parameters, in turn, considered the memorization strategies, the evolution of use, the adherence of graphical passwords to the suggested theme and the level of user satisfaction.

Users were divided into gender: male and female; and age: below thirty years old, between thirty and forty-five years old and above forty-five years old, as shown in Figure 4. Such division was necessary once the memorization factor was to be measured, which implied the need of considering different age groups.

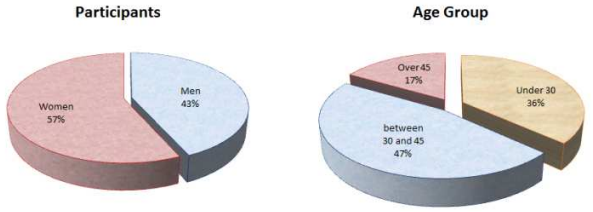


Figure 4 - Test users’ selection

Since one of the main criteria to be evaluated would be the capacity of memorization of passwords after a period of time, it was defined that the test would happen during an interval of fifteen days according to the schedule presented in Figure 5.

Figure 5 – Test schedule

V. RESULTS TABULATION

A. *Users behavior during icon exploration*

Initially, before the creation of passwords and beginning of the evaluation, the users were submitted to a process of exploring the icons from the first repertory in order to become acquainted with them, before the creation of the password itself. This process was performed for one minute, being twenty-five seconds for each category: people, objects, means of transportation and context/place.

With the support of the eye tracking tool it was possible to assess the icons that drew more attention during the exploratory process and the ones that were practically not observed, as shown in Figure 6 for the category “people”.

The Heat Map, a feature of eye tracking tool, enabled the visualization, throughout the variation of shades from green to red, of which icons were more observed, where the user fixed the attention and the icons that were practically not seen. The graph immediately after the Heat Map presents how many users choose each icon of the category. The red

 ISBN: 978-1-61208-229-5

"X" sign in the icons represents the ones that were not selected by the users when creating the password.

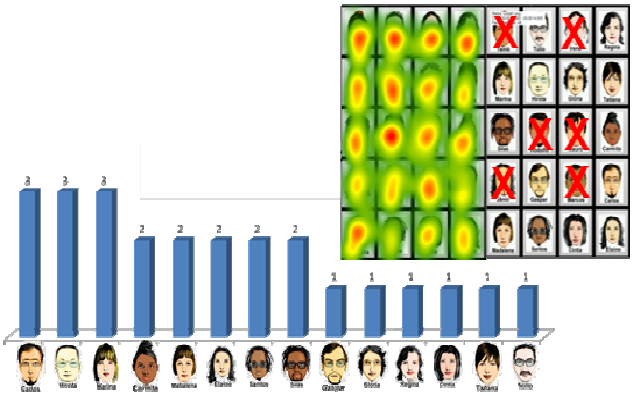


Figure 6 - Users behavior during the exploration of category "people"

B. Users behavior during the creation of passwords

The users selected to create passwords freely presented the behavior mapped in Figure 7.

Figure 7 - Users behavior in the creation of free passwords

According to the Heat Map mapping, it was possible to assert that the most observed icons where indeed chosen for the users passwords. Some icons had a high degree of choice in relation to others due to the easy connection to everyday situation and to cultural factors.

The users that had to choose the icons of the password according to a suggested theme also chose them freely and were not obliged to follow the theme. Since the test had the premise to create passwords that were safer and easier to memorize, the themes were intended to work as support in the creation of a plot for the composition of the password and ease of memorization. Figure 8 presents the passwords created with an associated theme and the adherence of each one of them to what was suggested.

From the ten users that participated in this category only four sought to use password adherent to the suggested themes.

Figure 8 - Users behavior in the creation of free passwords with suggested theme

C. Memorization strategies

Using of a repertory of icons separated into four categories, yet allowing for the formation of an episode with subject, actions and context, provided a great variety of memorization strategies and few participants declared the need to write down the password to help retain the icon set and at the end of the fourth use all participants were able to remember their graphical password in less than three attempts. These strategies were obtained by interviewing the users of the test after executing all interactions planned.

For all passwords used in this test, only eight users created a plot for the memorization process, six used the suggested theme and four affirmed not using any strategy.

The users that received the system generated passwords sought to look only at the icons given and not to the other icons in the screen to avoid possible confusions. For these users, the creation of a plot demanded a quick creativity which was not always possible.

Fig. 9 presents some passwords created by the users and the description of the strategies used, according to what was reported during the interviews.

Regarding password (I), the strategy used by the user was selecting icons of sports and food that he liked best. He chose the icon "balloon" because he wished to travel in one, the icon "airplane" because it was related to holidays and places he would like to visit.

Regarding password (II), the user admitted not using any strategy, only the visual memorization and chose the icon "calendar" because the date displayed was her husband's birthday. However, this user did not have a satisfactory performance during the interactions and made mistakes in different attempts.

 ISBN: 978-1-61208-229-5

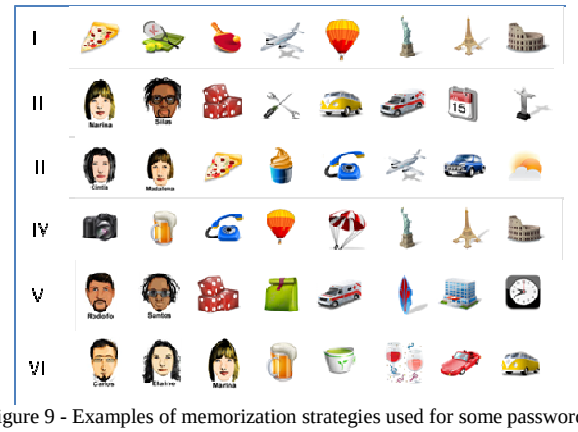


Figure 9 - Examples of memorization strategies used for some passwords created by the user

Regarding password (III), the user claimed to search for icons related to the suggested theme (romance), choosing outstanding pictures and remarkable or familiar names.

Regarding password (IV), the user chose objects he identified himself with and places he would like to visit.

Regarding password (V), the user also did not use any strategy, and since the password was provided by the system, he considered himself lucky by the presence of the icons "ambulance" and "hospital", in addition to two faces he could memorize by the names.

Regarding password (VI), the user claimed to have selected people with familiar names and icons related to two categories only (means of transportation and objects), thus avoiding confusion.

VI. RESULTS ANALYSIS

The results obtained in the data analysis, considering the performance during system interactions for all participants, were compared according to the group they belonged to, as follows:

- Participants with "User-Created passwords with No Training (UCNT)";
- Participants with "System-Generated passwords with No Training (SGNT)";
- Participants with "System-Suggested-Theme passwords with No Training (SSTNT)";
- Participants with "User-Created passwords With Training (UCWT)";
- Participants with "System-Generated passwords With Training (SGWT)";
- Participants with "System-Suggested-Theme passwords With Training (SSTWT)".

Figure 10 shows the average time of interaction after the first, fifth and ninth days from the creation date of the graphical passwords.

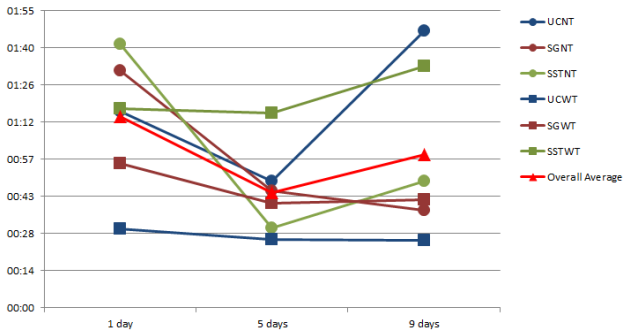


Figure 10 - Average time by user groups

As noted, the best time was achieved by the participants that created their own passwords and were trained (UCWT), while the worst time was achieved by the users that received a theme and were able to practice the password (SSTWT). It is possible to declare by the analysis of the graph in Figure 10 that four of six groups hold their performance close to the average, except for the last interaction for groups UCNT and TSCT, which is significantly deviated from the curve. It is also possible to observe that the groups that were not trained performed worse in the first day of use, but after the fifth day the performance returned to normal.

Regarding the creation method and training, a more detailed analysis done, to verify the behavior of users during the interactions. Figure 11 presents the average time of interaction of the participants, comparing passwords that were created by the users and the passwords generated by the system.

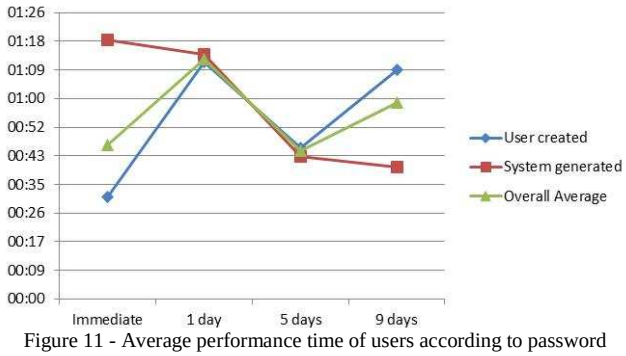


Figure 11 - Average performance time of users according to password creation method

The immediate use of system generated passwords had greater interaction time, as expected because the password was not created by the user and by the need to instantly having to create a mental model for its memorization. The same behavior was observed on the first day of use, after the password creation, where the time periods were considered long, and probably motivated by the user's fear of making mistakes in password selection. The behavior on the following days was considered better, curiously obtaining better performance on the last day for passwords that were not created by the user himself.

Figure 12 presents the behavior of the participants by previous training to the use of the created password. The

 ISBN: 978-1-61208-229-5

variable related to training was considered for this test as means to prove a possible outstanding performance regarding users that were not able to be trained.

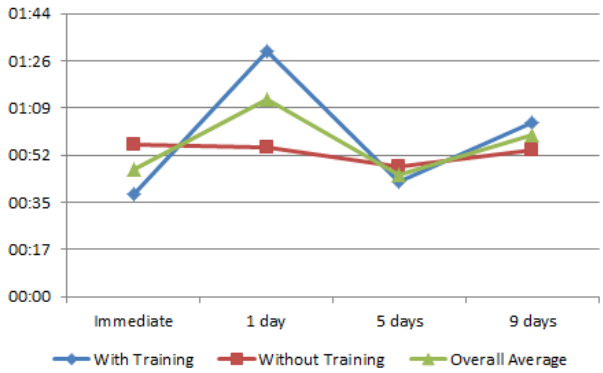


Figure 12 - Average performance of users with and without training

It was possible to observe in this study that only in the first day of interaction there was a significant difference between the participants with training and without training and, in an unexpected way, the users that were not trained had a better performance than the ones that were trained. This situation happened because many users did not use good memorizing strategies and didn't invested much time in creating the password, which means a lack of care during the creation process, making the memorization process difficult. Only after the first use that the memorizing strategy was created.

However, in the following days, the average for the participants with training and without training remained the same, which allows concluding that memorization is not directly related to the fact of password being practiced or not, but rather to the memorization strategy associated.

Regarding the experience in using graphical passwords, most of the users showed preference for this type of authentication, probably motivated by the innovation and ludic aspects of this approach. More than eight percent of the users considered the experience satisfactory and the remaining users since were not able to memorize the password in all interactions, eventually did not consider the experience as satisfactory.

It was possible to notice that in the last days, the degree of dissatisfaction increased in a significant manner, due to the interval between the previous interaction and the moment of password creation. Many memorization strategies here prove to be inefficient.

The same analysis was performed for the item related to the difficulty of use, where it was possible to observe a greater difficulty of the participants in the last day of interaction, linked to the time elapsed between the creation and penultimate day of use of the password; and also regarding to inefficient memorization strategies.

Since the memorization strategies did not work properly, the difficulty and quality of experience were compromised, not by the application and the solution itself, but often by the participant's own frustration for not accomplishing the task.

Regarding the category that most pleased the users there was a balanced result in categories "objects", "means of transportation" and "context/places" in the item most liked category.

The most significant result was related to category "people" with a larger number of critics mainly by the definition of the faces and confusion among several similar features. There were very few indifferent or totally satisfied users, not that significant to offset further analysis. Figure 13 shows the results obtained and the number of users that expressed their preference for any of the categories.

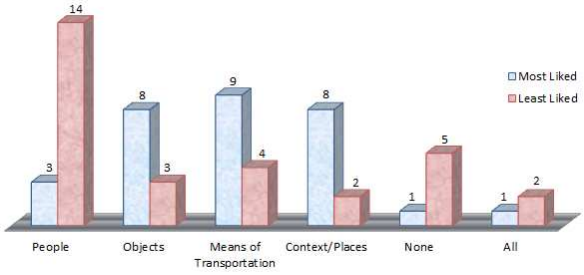


Figure 13 - Experience of use of graphical passwords

Regarding the difficulty of use, it was observed that less than ten percent of the users complained having problems using the prototype, being that this small percentage was composed by users that had no experience with smartphones, showing that the level of familiarity with technology created one of the main barriers in the usage of the proposed solution.

At last, regarding the tendency to replace alphanumeric passwords for graphical passwords, it was possible to observe a favorable result towards graphical passwords, but it was not so outstanding due to the fact of people being culturally adapted and proficient in the use alphanumeric passwords.

Even thou the fact of the experience with graphical passwords being interesting, from the tendency point of view, there is still some resistance, mainly because some considered this authentication method slower than alphanumeric passwords and identified a significant delay in the navigation between the icon screens.

VII. CONCLUSIONS

In the analyzes, it was possible to observe slightly better error rates and elimination of degenerated passwords through a more efficient repertory of icons with more graphically detailed images, greater possibility of creating stories and with fewer errors observed.

Regarding memorization, it was proved a reduction in the error rate and the users were surprised with their own performance, emphasizing that in the case of the category "people", it was easier to memorize the name of the person than the image itself, which many times was easily mistaken by the similarities of the faces presented.

The introduction of the process of training in the evaluation did not produced outstanding differences in the performance of the participants that had their passwords

 ISBN: 978-1-61208-229-5

generated by the system, but it affected in the creation of passwords by the users themselves, that obtained similar average input time and error rates.

Thus, the possibility of the creation of a script for memorization turns out not to be mandatory, since the less chosen icons where the ones with low contrast and difficult identification, while the ones most chosen had cultural appeal and high contrast.

From the usability requirements point of view, the decision to display icons in touchscreens was a good design solution however its performance needs to be improved. On the other hand, the feedbacks related to the quantity of icons selected and the navigation between screens ware not sufficiently clear and generated questions during the evaluation.

The size and quantity of icons available were well accepted by the users, requiring improvements in the process of cleaning the current selected icon which generated confusion and frustration in many users. The addition of an initial help mechanism is of extreme importance in a graphical authentication process, mainly by the paradigm change in authentication.

There was no correlation found between the most viewed icons in the exploration and the most chosen icons in the password composition. This suggests that the factors that made the icons more visually attractive and draw attention of the participants are not necessarily related to the criteria for choosing the icons to create stories.

Fewer users found great difficulties in the use of the prototype and the majority claimed to have a satisfactory experience of use during the tests. The adoption of this technology can be considered high, around sixty percent of the participants claimed they would replace their current alphanumeric passwords for graphical passwords, mainly for convenience of touchscreen and for the ludic experience involved.

The test conducted in the context of this project produced an analysis that is contributing for the definition of the graphical authentication solution to be used in the implementation of a functional prototype of BIOMODAL Project.

ACKNOWLEDGMENT

The authors acknowledge the financial support given to this work, under the project “Biometric Multimodal and Graphical Authentication for Mobile Devices – BIOMODAL”, granted by the Fund for Technological Development of Telecommunications – FUNTTEL – of the Brazilian Ministry of Communications, through Agreement Nr. 01.09.0627.00 with the Financier of Studies and Projects - FINEP / MCTI.

REFERENCES

[1] B. Kirkpatrick. “An experimental study of memory”. Psychological Review, 1894, 1:602-609.

[2] S. Madigan. “Picture memory”. In J. Yuille, editor, “Imagery, Memory, and Cognition: Essays in Honor of Allan Paivio”, cap.3, pp. 65-89. Lawrence Erlbaum Associates, 1983.

[3] A. Paivio, T. Rogers, and P.C. Smythe. “Why are pictures easier to recall than words?” Psychonomic Science, 1968, 11(4):137-138.

[4] R. Shepard. “Recognition memory for words, sentences, and pictures”. Journal of Verbal Learning and Verbal Behavior, 1967, 6: pp. 156-163.

[5] Passfaces Corporation. “The Science Behind Passfaces”. Write paper, http://www.realuser.com/enterprise/resources/ whitepapers. htm, accessed Sep, 2012.

[6] R. Dhamija and A. Perrig. “Déjà Vu: A user study using imagens for authentication”. In 9th USENIX Security Symposium, 2000. Proceeding of the SSYM'00 9th conference on USENIX Security Symposium - Volume 9, Pages 4 – 4. USENIX Association Berkeley, CA, USA.

[7] R. Biddle, S. Chiasson, and P.C. Van Oorschot. 2012. “Graphical passwords: Learning from the first twelve years”. ACM Comput. Surv. 44, 4, Article 19 (September 2012), 41 pages.

[8] F. Monrose and M. Reiter. 2005. “Graphical passwords. In Security and Usability: Designing Secure Systems That People Can Use”. L. Cranor and S. Garfinkel, Eds. O'Reilly Media, Stebastopol, CA, Chapter 9, 157-174.

[9] S. Xiaoyuan, Y. Zhu, and G. Scott. Owen. 2005. “Graphical Passwords: A Survey”. In Proceedings of the 21st Annual Computer Security Applications Conference (ACSAC '05). IEEE Computer Society, Washington, DC, USA, 463-472.

[10] K. V. Renaud. 2009. “Guidelines for designing graphical authentication mechanism interfaces”. Int. J. Inf. Compute Security 3, 1 (June 2009), 60-85.

[11] A. De Angeli, L. Coventry, G. Johnson, and K. Renaud. 2005. “Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems”. Int. J. Hum.-Comput. Stud. 63, 1-2 (July 2005), 128-152.

[12] D. Davis, F. Monrose, and M. K. Reiter. 2004. “On user choice in graphical password schemes”. In Proceedings of the 13th conference on USENIX Security Symposium - Volume 13 (SSYM'04), Vol. 13. USENIX Association, Berkeley, CA, USA, 11-11.

[13] C. A. Tambascia, E. M. Menezes, R. E. Duarte. “Usability Evaluation Using Eye Tracking for Iconographic Authentication on Mobile Devices”. Mobility 2011, The First International Conference on Mobile Services, Resources, and Users. Barcelona, Spain (October 2011), 117-122.

[14] P. Dunphy, A. P. Heiner, and N. Asokan. 2010. “A closer look at recognition-based graphical passwords on mobile devices”. In Proceedings of the Sixth Symposium on Usable Privacy and Security (SOUPS '10). ACM, New York, NY, USA, , Article 3 , 12 pages.

[15] Passfaces Corporation. “The science behind Passfaces”. White paper, http://www.passfaces.com/enterprise/resources/ whitepapers.htm, accessed September 2012.

[16] I. Avila, E. Menezes, A. Braga. “Memorization Strategy for Iconic Passwords”. In: IADIS International Conference Interfaces and Human Computer Interaction 2012, Lisboa. Proceedings of the IADIS Intl. Conf. Interfaces and Human Computer Interaction 2012. Lisboa: IADIS, 2012. v. 1. p. 123-132. 1.

 ISBN: 978-1-61208-229-5

Socio-technical Study of Teleworking:

From the analysis of employees'uses to the design of a virtualized and unified platform

Valérie Fernandez
Department of Human Sciences
Paris Institute of Technology – Telecom ParisTech
Paris, France
Valerie.fernandez@telecom-paristech.fr

Laurie Marrauld
Department of Human Sciences
Paris Institute of Technology – Telecom ParisTech
Paris, France
Laurie.marrauld@telecom-paristech.fr

***Abstract*— In this working paper, we present the project "WITE 2.0". This project is at the crossroads of various issues related to mobility (mobility turn) and use of Information and Communication Technologies. Wite 2.0 is a part of the designing process of a collaborative communication tool: "a virtualized and unified platform". We define scenarios of teleworking practices, "equipped" by ICTs, and use these scenarios to better specify the platform. The project started at the end of 2010 for a period of 18 months. The analysis is based on several complementary methodologies: a qualitative study (47 semi-structured interviews), a quantitative survey (553 respondents), and an experimentation of the platform. In this paper, we present the main results of the interview survey through the following themes: remote management, skills, articulation of private and professional spheres, and the maturity of technologies. We also describe how these elements help to understand the workers' practices evolutions.**

Keywords: teleworking; ICT; management; socialization; competences.

I. INTRODUCTION

The project WITE 2.0 (Work IT Easy) [10] is a research and innovation program supported by public funds. It is a multi-partner project (academic and industrial actors).

The aim of this project is to create a virtualized platform. This platform represents a unified work environment, based on virtualization, instant communication and interoperability of systems, and it allows the individuals to work anywhere (and possibly anytime).

The platform is a software solution that centralizes the access to a set of functionalities, originally offered by several applications: it's the principle of unified communications. Unified communications (UC) is the integration of real-time communication services such as instant messaging (chat), presence information, telephony (including IP telephony), video conferencing, data sharing (including web connected electronic whiteboards aka IWB's or Interactive White Boards), call control and speech recognition with non-real-time communication services such as unified messaging (integrated voicemail, e-mail, SMS and fax).

This platform is accessible from any connected terminal, either fixed or mobile (desktop, laptop, tablet, Smartphone, etc.). The WITE 2.0 platform will provide a wide range of communication tools that can be activated on demand in different situations, and depending on users' needs (VoIP, discussion groups, instant messaging, email, etc.).

The project has four main stages, divided into several subsections each. It is supported by a socio-technical analysis. Telecom ParisTech has assumed leadership in the scientific study of the needs and uses by administrating semi-structured interviews with individuals regularly working "remotely". We wanted to better characterize these work situations: at home, on the premises of the employer but in geographically dispersed locations; in telecentres / co-working spaces / business center, with geo-distributed teams working.

The rest of this paper is organized as follows. In the Section II, we describe the different stages of WITE 2.0 project; we specify the notion of virtualization. In Section III, we expose the socio-technical study with the chosen methodology and the main results. We conclude by giving some research prospectives.

II. DESCRIPTION OF WITE 2.0. PROJECT

A. The project issues

Mobility at work is spreading in the context of the mobility paradigm's evolution [9]. The project WITE 2.0 intends to address the urgent need for solutions in the field of remote collaborative work. These needs include ways to collaborate, communicate and socialize, but also to access these features regardless of the location, and from any workstation. It will provide a unified interface integrating all the features, and having a wide range of communication tools selectable on demand.

B. The project stages

The project WITE2.0 is divided into four major phases. The first one concerns the study of employees' needs and uses, for remote collaborative work. In order to capture the needs and uses, we conducted two surveys: first, a qualitative study based on 47 semi-structured interviews and second, an online quantitative survey with 553 individuals. The object of this phase is to highlight the kinds of remote collaboration in order to make recommendations related to the design of the platform. The results are published in the report "Work, socialize and collaborate remotely" [5]. The

 ISBN: 978-1-61208-229-5

main preliminary results of this report focus on how to socialize in a context of teleworking, the question of remote management, and of the technical skills needed for the use of ICT. The recommendations focus on the access to digital resources, the business information systems, and on issues related to security.

The second phase of the project focuses on the technology. It is divided into two parts. The first part consists of the writing of functional and technical specifications of the platform WITE2.0. This document contains descriptions of service needs, and a comparison of various existing virtualization solutions. We notice that, since the launching of the project WITE 2.0, some other virtualization solutions have appeared on the market (Citrix, etc.) [10]. Through the comparison of different solutions, we highlight the distinguishing features and the technological services of the platform WITE 2.0. The second part of the phase 2 includes the implementation of all the technical elements necessary for the platform WITE2.0. These technical elements include the virtualization, the development of a socialization software solution, the services integration, the development of a unified software, and a beta testing of the platform.

In the third phase, the project partners have introduced new technological elements for the components and voice applications, the mobile profiles, and the SIP recorder.

C. The virtualized workstation

The major technical element of this platform is based on the virtualization of information systems (IS). The workstation virtualization solution that we are interested in is also called the "PC on demand."

The virtual workstation *displays* a virtual image on the user workstation that is *executed* on a remote server (not virtualized). This technology has several advantages (the list is not exhaustive):

- Centralizing logical components i.e. the operating system (Windows or Macintosh) but also applications (such as Word, PowerPoint, Skype);
- Checking the lifecycle of workstation: One can decide to create or delete virtual PC on demand);
- Managing storage resources: as the main storage element is no longer the physical position of the hard drive on which the employee works;
- Access to a virtual portal: to recreate a selection of applications (i.e. a library of applications) each time the PC is created, and customized for each company;
- Access to an individual virtual PC guaranteeing better mobility management of the employee (the employee creates and destroys the PC to the demand for mobility).

Virtual workstations will address a number of challenges compared to “ordinary” workstations (especially at the administrative, security and deployment of machines levels). Virtual machines, for example, can decrease functional costs (maintenance, etc.), and technical problems such as obsolescence of the workstation. With VDI architecture, the ISD has no longer constraints related to maintenance and administration of its fleet of workstations. The user is no longer dependent on a single physical computer and can connect to his "own PC" from different physical devices, even from terminals like thin clients.

Hence, virtualization is an already existing technology. The value of the WITE 2.0 platform is to combine this exiting technology with a collaborative tool (unified communication). The aim is actually double: in one hand, to reduce (or eliminate) the problems due to data security, and on the other hand, to improve the level of collaboration between employees.

III. A SOCIO-TECHNICAL STUDY

A. What about “ remote collaborative work” ?

In a context of managerial culture based on “face to face”, the new paradigm of mobile work does not seem to establish in French companies. Some forms of work organization are deeply rooted. Some managers and IT system directors still reluctant to introduce new developed technologies because of the security of the data circulation. Our bias is to say that the paradigm shift can take place now if we take into account the issues of teleworking in "sociological" and "technical" terms: hence, the importance of analyzing the practices of work organization and use of ICT, and also the experimentation of the virtual platform.

B. What is teleworking ?

In our analysis, we are interested in a key notion: the concept of "teleworking" that we have considered in its most classic form (the homework), but also in the most diverse realities that it could be today: either, all forms of "remote working", i.e., forms of organization and / or performing work outside the classical unity of time and place. Indeed, many studies emphasize that the unity of time and place that characterized the traditional organization of work, would tend to disappear [2][3][4][6].

Thus, the definition of teleworking that we have selected is based on:

- The fixed place of work or alternating between several workplaces, provided they are removed from the hierarchy and/or colleagues;
- The relationship to the employer and colleagues, remotely and by electronic links, thus justifying the name of teleworking [8].

C. Methodology

The first results presented in this paper are based on a qualitative analysis approach. We have particularly studied the practices of coordination and cooperation in various configurations of remote work, more specifically in management practices supported by different communication technologies (fixed or mobile). We believe this kind of qualitative study is the most relevant because we make a statement about teleworkers’ practices. As the virtualized

 ISBN: 978-1-61208-229-5

and unified platform technology is designing, we have realized that we actually had little knowledge about the current technologies' practices' realities. The classic typology of the four kinds of teleworkers – homeworkers, mobile workers, telecenter worker, virtual team worker [1] - should really be evolved. We have decided to question the realities of the teleworkers'practices to understand the evolution of the work organization et to link the technology to specific uses (or link uses to specific technology). Hence, we have thought that the most equipped teleworker will be the most graduated (with the most responsabilities). We have discovered that the uses of mobile ICTs evolved in a very paradoxal way.

During the operational phase of the study, from May to July 2011, 47 interviews were conducted face to face with workers performing work remotely. Our sample of interviewees was compiled from relays (managers and human ressources services of companies).

These interviews, lasting an average of 1:30 each, were the subject of subsequent detailed accounts.

The sample is constituted with different profiles of employees. In our analysis, we specially separated three kinds of profiles, regarding the level of the management's activity:

- The entrepreneurs: they are supposed to be very autonomous in their work, self-managed, and often have their own company.
- The executives and intellectual professions: Graduated employees (engineers, etc.), they always have employees to manage. Relatively autonomous, they sometimes have to report their activity.
- The associate professionals and employees: Executives manage them. They have technical activities.

Spreadsheet 1: Sample

	Homework	**Alternating homework**	**Mobile worker**
Entrepreneurs	2	2	2
Executives and intellectual professions	1	17	14
Associate professionals - Employees	2	4	3
Total	**5**	**23**	**19**
Women	3	8	4
Men	2	15	15

D. First results of the analysis

The first results of the analysis cover various aspects of remote working of the surveyed employees:

The articulation of private and professional life. Moreover, a phenomenon seems banal, the activity overflow telecommuting seem becoming more and more important.This form of teleworking involves a regular work done outside of the official working hours, most often at home. These activities are mostly related to checking and reply to emails. People build tactics to separate private and professional spheres. They are based, for example, on partial *reachability* over the mobile phone. People choose to disconnect and not to check their telephones after a defined period. This strategy is the most intuitive. But, some people choose to disconnect their mobile phone to work at home without being disturbed every time. Some of the interviewed said that it is difficult to concentrate on doing a task when:

- they are at the office (colleagues interrupt them all the time, or there is lot of noise, etc),
- they are at home when the mobile phone is switched on (or even when just the Internet is connected).

Forms of remote "socialization": social networking tools seem to be effective. This proximity can develop communicative relationships of trust, explaining in part the stability of the worker community [7]. For example, the unified communication and the social media, that employees use, seem to be very convenient. They use to talk to each other through the *chatrooms* without ever see each other in real life. They become friends in other social media (Facebook, etc.) and they develop a real relationship of complicity and friendship. It's difficult to say if this relation is also based on trust.

New forms of remote management: ICT are, for certain categories of workers, as a digital control infrastructure, replacing the physical presence of the manager (control connection time, obligation of permanent reachability via the mail or instant messaging). The current technology controls very well the activity of the employees. The managers can hear the phone calls between employees and clients without the employee's knowledge. In this case, the statut of presence and the activity on the chat are essential to proove that the employee is working.

In other cases, some managers construct tactics of "motivation" by using ICTs as a lever of the collective dynamics (*Chat* between collegues).

Skills development: It appears that many people working remotely learn to use ICT more or less on a mode "self-taught", "on the job". Some of them practice forms of "tinkering" computer, as the diversion of scripts use of some business applications, etc.

Analysis of current technology: current technology does not allow to maintain conditions comparable to those offered in the premises of the employer. Sometimes, these conditions are not reproducible. In this case, the employee is forced to fragment his work activity. He will assign, for example, specific tasks to places where he works according to the possibilities offered by his work environment. Furthermore, this segmentation may come from the desire to choose a specific work environment, quieter and less disruptive than in the enterprise to perform tasks requiring more concentration.

 ISBN: 978-1-61208-229-5

Based on the results of the qualitative study, we suggest some recommandations for the WITE 2.0 platform:

- Firstly, the WITE 2.0 platform must be integrated in the IS of the companies: a lot of interviewed workers pratice some “tinkering” because of the restrictions of the IS of their company. Workers want to access to the tools that they need, without having problems with the technical services.

- Secondly, we suggest adding some filters to specify the time of reachbility and the time of disponibility. A lot of workers are often disturbed by their colleagues during their activity; because the colleagues do not check if they are busy or not (at the office and, also, at home by ICT). These filters could be used for teleworking and for presentiel working (“I check if my college is available before disturbing him in his office”).

- Finally, we suggest to labeling the information flows received everyday. We propose to label these flows in function of the emergency of the data, the nature of the data, the working group which is concerned, and the person who send the data, etc. For example, email box could be integrated to the social media as a secondary function and people could juggle with different media (chat, email, timeline, etc) subject to their needs.

IV. DISCUSSION AND CONCLUSION

There were several configurations of "remote working" that we observed that changed our usual representations of the "teleworker." We propose here three “stereotype” representations.

First, the figure of the teleworker, exclusively at home, whose teleoperator is an ideal-typical figure. He uses a desk in a corner of the room or the bedroom, not in a room dedicated solely for homework. However, the boundaries between private and professional life are maintained due to strong control of his activity by the ICT. In fact, ICT is, for him, a "control infrastructure". Despite of the physical distance, hierarchy is near: hours are controllable via the use of ICT, including instant messaging to verify that the teleoperator is well behind his computer. The remote control via ICT is like a substitute of direct managerial control.

Second, the figure of the mobile worker who sometimes works at home. ICTs provide a permanent reachability that does not always correspond to their availability. Technologies act as strong regulators of space-time job. On the contrary, the executives act more like disrupting their working space in confusing the virtual presence status (reachable / available). Sometimes, the mobile worker can be considered as a “techie” worker, clamped in his uses of ICT.

Third, the figure of the worker in a co-working space or in a telecenter: with a high degree of autonomy in organizing his work, the worker uses this structure to "frame" his activity (immerse himself in a group is a way to put boundaries between private and professional spheres), but also to densify his socioprofessional network IRL (In Real Life). The worker in a co-working space will tend to regulate his working hours, helping to build boundaries between private and professional spheres. He can also use the dynamic shared places to gain in competence, although it is usually relatively independent with technology (although this may evolve with the growing phenomenon of collaborative workspaces).

The precise definition of teleworkers is difficult to establish, as the profiles are varied and the situations are diverse (new working configurations including practices of re-sedentarization activity). The characteristics of these new positions (the nature of risks, for instance) question both at the managerial level (the hierarchy acceptance) and technical level (the ISD acceptance). They also question the employees themselves who sometimes try to make self-regulation. Embedded in a double paradigm (technological and organizational), teleworkers seem to evolve in unexpected ways, even paradoxical. In order to understand the realities of working remotely, one is asked to investigate the tools for teleworking in their collective aspect of interaction management, the security of sharing data and the practices of professional social tools. The technical characteristics of these tools are revealed only through the "remote" collective dynamics and vice versa.

REFERENCES

[1] T. Breton, (1996) “ Le télétravail en France : Situation actuelle, perspectives de développement et aspects juridiques ”, Rapport au Ministre d’Etat, Ministre de l’Intérieur et de l’Aménagement du Territoire et au Ministre des Entreprises et du Développement Economiques, La documentation française, Paris. English translation “Teleworking in France: Current Situation, development perspectives and juridic aspects” *Report to the prime minister*.

[2] L. Chen, and R. Nath, “Nomadic culture: cultural support for working anytime, anywhere”, Information Systems Management, Fall, pp. 56-64, 2005.

[3] F. Cocula, and A. Fredy-Planchot, “Pratiquer le management à distance ”, Gestion 2000, n° 1, pp. 43-63, 2003.

[4] G.B. Davis, “Anytime/anyplace computing and the future of knowledge work”, Communications of the ACM, vol. 45, n° 12, pp. 67-73, 2002.

[5] V. Fernandez, C. Guillot et L. Marrauld “Travailler, se sociabiliser et collaborer à distance”, rapport de recherche Wite 2.0, English translation “Work, socialize and collaborate remotely”, research report, 2011.

[6] M.L. Gribbins, J. Gebbauer, and M.J. Shaw, “Wireless B2B mobile commerce: a study on the usability, acceptance, and process fit”, Ninth Americas Conference on Information Systems, Tampa Florida USA, August 4-6, 2003.

[7] J. Rosanvallon, “Travail à distance et représentations du collectif de travail ”, Revue Interventions économiques, 34 | 2006, posted online 07/01/2006, Accessed on 09/17/2012. URL : http://interventionseconomiques.revues.org/706.

[8] L. Thomsin, *Télétravail et mobilités*, Les Editions de l’Université de Liège, Collection « Synopsis », 2002.

[9] J. Urry, *Mobilities*, Cambridge, UK: Polity, 2007.

[10] J., Wang, L., Yang, M., Yu, S., Wang, “Application of Server Virtualization Technology Based on Citrix XenServer in the Information Center of the Public Security Bureau and Fire Service Department”, International Symposium on Computer Science and Society, IEEE conference publications, pp. 200-202, Kota Kinabalu, Malaysia, 16-17 July, 2011.

[11] Web Site WITE 2.0 project : http://www.wite2-0.fr/ Accessed on 09/17/2012.

 ISBN: 978-1-61208-229-5

User Acceptance and Usage Continuance of Interactivity Enabling Technologies

A mixed method approach to the evaluation of acceptance and usage continuance of NFC applications

Elisabeth Pergler, Verena Skedel
evolaris next level GmbH
Graz, Austria
elisabeth.pergler@evolaris.net, verena.skedel@evolaris.net

Abstract— **This paper presents the results of an exploratory study concerning user acceptance and usage continuance in the field of interactivity enabling technologies. Participants had the chance to try Near Field Communication (NFC) technology in four different usage scenarios and thereby assuming different specified roles. In the course of this usage experience, quantitative data was collected by means of traditional standardized acceptance research instruments (technology acceptance model, unified theory of acceptance and use of technology and expectation disconfirmation theory) and qualitative data was gathered in form of user comments. The data was then compared using a mixed method approach in order to find out whether traditional instruments are applicable to acceptance research of interactivity enabling technologies such as NFC. Our results show that applying traditional instruments will cause a significant loss of valuable information and the results are of limited relevance for the design of specific applications. It is, therefore, concluded as the main output of this paper that future acceptance research in this field will need to include qualitative data, but, at the same time enable collection of huge numbers of user opinions as standardized quantitative methods will provide.**

Keywords-technology acceptance; interactivity enabling technology; mixed method; TAM; UTAUT; expectation disconfirmation theory;

I. INTRODUCTION

Technology acceptance research is a crucial task in the development process of mobile applications. Acceptance is regarded as the adoption of a new technology and its further usage as many business models in this field are based on repeated usage. Specific characteristics of mobile devices enable their usage in highly dynamic contexts [1], which require dynamic methods of acceptance research. Traditional acceptance research might not be appropriate for this dynamic task as prior research did indicate shortcomings in the area of mobile technologies in general [2]-[4]. Many new applications are based on interactivity enabling technologies such as NFC. In the context of this research project, interactivity enabling technologies are defined as technologies that support or enable interaction between humans and objects or among humans by means of mobile devices. NFC is only one example of such an enabling technology. These technologies also require acceptance evaluation, but might not be assessed by traditional methods of acceptance research as the technology itself is not a perceivable characteristic of an application or service but acts as an enabler for it. The user might, therefore, not even be aware of the technology, which is the basis for the service or application. Nevertheless, it is inevitable to find out, which enabling technologies are acceptable, and which are not. The main research question in this paper is, therefore: How can acceptance and usage continuance of interactivity enabling technologies are assessed?

The research questions in detail are:

- Will application of traditional instruments of acceptance measurement provide useful information in the context of interactivity enabler technologies such as NFC?
- Are there similarities between acceptance factors that are observed by means of qualitative research and those measured by traditional acceptance instruments?

By addressing these questions in an exploratory study, it is intended to uncover potentials for future research in the area of interactivity enabling technologies and to gain a better understanding of unique characteristics of these technologies, which affect acceptance. In order to achieve this goal, user tests were conducted and several traditional commonly used instruments were applied as well as qualitative methods of data gathering in a mixed method setting. The comparison of the obtained results is the core issue of this paper. The remainder of this paper is organized as follows. In Section 2, the state of the art acceptance models are discussed together with commonly used methods of technology acceptance research. The methodology that was used for empirical testing is presented in Section 3 and results are provided and discussed in Section 4 followed by concluding remarks and an outlook on future research activities and questions.

II. STATE OF THE ART

The most often used model in technology acceptance research is Technology Acceptance Model (TAM), which explains acceptance by means of two key factors [5]:

- Perceived ease of use
- Perceived usefulness

Almost half of all papers in the area of mobile technology acceptance are based on this model [6]. Prior research did show that application of TAM might lead to inconsistent results. This is why process theories are recommended that include experience/feedback loops [7] and [8]. They enable researchers to capture dynamic

 ISBN: 978-1-61208-229-5

processes and interaction between technological and organizational structures. In the original study [5] a follow up two weeks after initial data collection did also indicate significant changes of user perceptions over time.

Unified Theory of Acceptance and Use of Technology (UTAUT) is a compound model that includes elements of TAM and seven other models [9]. Among these models are motivational, social cognitive and diffusion models. The constructs included in UTAUT are:

- Performance expectancy
- Effort expectancy
- Attitude towards using technology
- Social influence
- Facilitating conditions
- Self-efficacy
- Anxiety
- Gender, age, experience and voluntariness of use as moderators
- Behavioral intention

Expectation-disconfirmation theory [10] is used in the field of technology acceptance research in order to "move from traditional static IT usage models (e.g., TAM, TPB, TAM2) to temporal models focusing on understanding fluctuation patterns of IT usage." [11]. The theory has been applied on TAM and data were gathered ex post [12] or as a two-stage research design where expectations are captured before usage and confirmation or disconfirmation after hands-on experience [13]. Three-stage designs were used to show that expectations will experience stabilization and become more consonant with experience after longer periods. This kind of research design includes different constructs in the questionnaires at three points in time (t1, t2, t3) [11]:

- Usefulness (t1, t2, t3)
- Attitude (t1, t2, t3)
- Disconfirmation (t2, t3)
- Satisfaction (t2, t3)
- Intention (t2, t3)

Many data collection methods for dynamic capturing of user behavior limit the number of possible data sets as they are time-consuming and laborious. This is especially true for shadowing where the researcher follows the user in the field and observes and documents the user behavior. In addition to a high expenditure of time it is also probable that the user is disturbed by the researcher in his natural environment [14]. A similar case is contextual field research where ethnographers capture user activities by means of photos and communication sequences and combine them with context information on a time line [15]. User generated content enables the collection of numerous user opinions that can be analyzed quantitatively or qualitatively. Most distribution platforms of mobile applications, e.g., Apples AppStore, include user generated content in form of user reviews. These text documents benefit from the voluntariness of their provision in contrast to questionnaire-based surveys, which limit the range of possible answers by standardization [16]. Another non-reactive method is behavior tracking, where user simulations are computed [17] in order to simulate for instance minimum requirements of service quality [18] or to document user mobility behavior [19]. Automated event protocols, however, disregard motives and causes of user behavior to a large extent.

III. METHODOLOGY

An exploratory approach was chosen in order to obtain valuable insights regarding actual user acceptance factors. Therefore, the study was designed according to state of the art methods and instruments of technology acceptance research and also includes further qualitative measures.

A. *Research Design*

Trialability is an important factor of technology acceptance measurement [20]. This is why there should be hands-on experience included in the test setting. Questionnaire-based surveys that rely on mere imagination of technologies, which the participant never used himself are not as valid as those conducted after usage experience though previous studies indicated that pre-prototype usefulness measures are able to approximate usefulness measures after hands-on experience quite well [21].

There are good reasons for field studies as well as in favor of lab studies. In the context of emotion capturing it is common to prefer field studies because users should experience technology in normal usage situations and emotions may be different in artificial laboratory setups [22]. Especially mobility as a key characteristic of mobile technologies is hard to simulate in lab studies. Nevertheless most evaluations of mobile systems are designed as laboratory tests [1]. Laboratory studies are preferred in cases that require experimental control of unknown variables and they simplify data collection [1]. For evaluation of product characteristics lab tests are commonly regarded to be sufficient [23]. We decided on an experimental approach as NFC is not a widely used technology yet. Applications are rare in the field and need to be tested in a lab environment. In order to reduce disadvantages of lab studies and to foster imagination of NFC application opportunities the test setting was designed as role plays. Two participants were interacting in predefined roles that were close to reality and in that role experienced different interactive NFC applications.

There were 30 participants of which 14 were male and 16 female. The age of the participants ranged from 20 to 40 and was 28.03 years on average with standard deviation of 4.97. Concerning the general attitude regarding new technology the sample appeared to be rather technology affine as depicted in Figure 1.

 ISBN: 978-1-61208-229-5

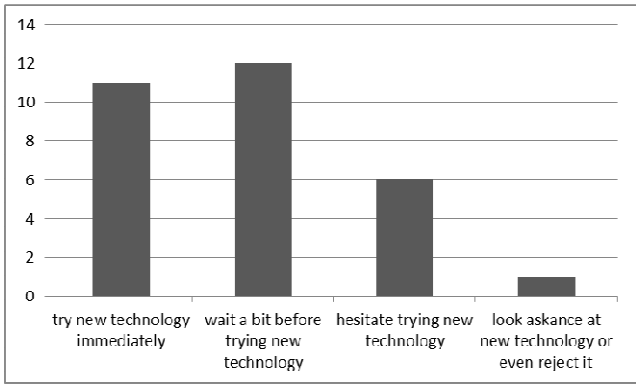


Figure 1. Technology affinity of the participants regarding their overall attitude towards new media and technologies.

B. Schedule

When the participants arrived at the test site they were introduced to interactive technologies by the example of NFC. A brief description of the technology was performed as oral presentation by the interviewer. Immediately after this introduction the participants were asked to fill in the first questionnaire concerning their expectations regarding NFC. The participants were then assigned to their roles.

In the course of the role plays, there were four tasks to perform:

- 1st task – user of social media

Imagine you are a user of Facebook. You have a Facebook account and have already updated your status and checked in at different places before. The name of your account is "Evo Laris". Further imagine you just came to our company and you would like to capture your first visit of the TecLab (laboratory with technical equipment where user tests usually take place) in form of an update status and check in.

Technical equipment: one NFC tag attached to a plain surface on the entrance door that initiates a status update and a check in when the NFC enabled mobile phone is within activation distance

- 2nd task – business meeting

Imagine you are person A (fictitious name equivalent to John Q. Public, in the following referred to as just "person A") and you meet person B (fictitious name, in the following referred to as just "person B") for business related reasons. Person B is a potential customer of person A and you never met before. After settling the details of a contract you want to interactively exchange contact information (name, phone number, email address, postal address) for further proceeding. In order to do this you can use the mobile phone.

Technical equipment: business cards for person A and person B including NFC tags that initiate the inclusion of contact data into the address book of the mobile phone when the NFC enabled mobile phone is within activation distance

- 3rd task – customer of a retailer for consumer electronics

You just bought a flat screen TV set. The device was delivered at your home and you set it into operation successfully. Your friend person B is interested in technical details concerning the device and you cannot find the manual at the moment. As the technical key points don't come to your mind immediately and the device is not self-explaining you want to get further information. In order to get this information you can use the mobile phone.

Technical equipment: one flat screen TV set with an NFC tag that is attached to its surface and initiates download of the manual when the NFC enabled mobile phone is within activation distance

- 4th task – participant of a fair

Imagine you participated in a congress. After you, person A, and your colleague, person B, entered the fair area you want to orientate yourselves. You want to get an overview of companies' display booths and their locations. Moreover you want to know who is going to present which topic and when. In order to get this information you can use the mobile phone.

Technical equipment: one NFC tag that that is attached to a plate with the conference name and NFC logo on it and initiates download of a congress program when the NFC enabled mobile phone is within activation distance

The necessary equipment for all four tasks was prepared in the TecLab and the participants were provided with NFC enabled mobile phones (Samsung Nexus S). As most participants were not familiar with the usage of this specific device they received a brief instruction to the handling and functionalities. During the tasks the interviewer took notes concerning observational data (duration of task accomplishment, did the participants try to solve the problems together, participant reactions) and some open question data (difficulty of the task, suitability of NFC for the specific situation, comments and suggestions for improvement).

Following to the tasks, the participants had to fill in the second questionnaire concerning their experiences with NFC technology.

C. Research Instruments

We used a mixed method approach including quantitative data from standardized questionnaires and qualitative data from user comments. Research instruments were adapted from traditional technology acceptance instruments. The first questionnaire was based on expectation disconfirmation theory [11] except for the construct perceived usefulness for which we used all the original TAM items [5].

The second questionnaire was based on TAM [5], UTAUT [9] and expectation-disconfirmation theory [11] and [24]. Several items of the performance expectancy scale were excluded because of redundancy with other scales and the scale for behavioral intention was also reduced as the different meanings of the three expressions were not translatable into German language. The items that were used are listed below:

Perceived usefulness [5]:

- All items from the original instrument.

Perceived ease of use [5]

- All items from the original instrument.

Performance expectancy [9]:

 ISBN: 978-1-61208-229-5

- If I use NFC, I will increase my chances of getting a raise.

Attitude toward using technology [9]:

- All items from the original instrument.

Social influence [9]:

- All items from the original instrument.

Facilitating conditions [9]:

- All items from the original instrument.

Self-efficacy [9]:

- All items from the original instrument.

Anxiety [9]:

- All items from the original instrument.

Behavioural intention to use the system [9]:

- I intend to use NFC in the next <n> months.

Disconfirmation [11]:

- All items from the original instrument.

Satisfaction [11]:

- All items from the original instrument.

Attitude [11]:

- All items from the original instrument.

Intention [24]:

- I intend to continue using NFC rather than discontinue its use.
- My intentions are to continue using NFC than use any alternative means.
- If I could, I would like to discontinue my use of NFC.

In addition, the time span needed for task completion was documented as well as spontaneous reactions of the participants during the scenarios. The participants also had to grade the appropriateness of NFC technology for the specific task as well as the difficulty level of task completion on a scale ranging from 1 – very good to 5 - poor. After filling in the questionnaires they were asked to comment on the used research instruments.

IV. RESULTS AND DISCUSSION

We computed correlations among all constructs included in our questionnaire in order to find out, which constructs influence behavioral intention the most. The first step was the computation of mean values corresponding standard deviations for all constructs are listed in Table 1.

TABLE I. MEANS AND STANDARD DEVIATIONS OF ALL TESTED CONSTRUCTS

	Mean	Standard deviation	N
Disconfirmation (DIt2)	3.63	0.77	29
Perceived usefulness (PUt1)	4.43	1.63	30
Attitude (ATt1)	3.90	0.79	30
Perceived usefulness (PUt2)	3.96	1.92	30
Perceived ease of use (PEt2)	1.94	1.13	30
Attitude toward using technology (AUt2)	3.92	0.79	30
Social influence (SIt2)	2.43	1.08	29
Facilitating conditions (FCt2)	3.16	0.77	30
Self-efficacy (SEt2)	3.28	1.02	30
Anxiety (AXt2)	1.74	0.69	30
Satisfaction (SAt2)	3.96	0.64	30
Attitude (ATt2)	4.10	0.88	30
Intention (INt2)	3.82	0.91	30

TABLE II. CORRELATIONS AMONG CONSTRUCTS, N=29, *P<.05; **P<.01

	DIt2	PUt1	ATt1	PUt2	PEt2	AUt2	SIt2	FCt2	SEt2	AXt2	SAt2	ATt2
DIt2	1											
PUt1	-.35	1										
ATt1	.09	-.11	1									
PUt2	-.69**	.80**	.01	1								
PEt2	-.33	.10	-.07	.25	1							
AUt2	.74**	-.55**	.37*	-.78**	-.19	1						
SIt2	-.07	-.22	.16	-.18	-.37	0.13	1					
FCt2	-.32	-.04	-.20	.09	-.17	-.26	.36	1				
SEt2	.35	-.24	.04	-.28	-.31	.26	-.04	-.12	1			
AXt2	-.01	.05	.03	-.13	-.08	.11	.03	-.24	-.00	1		
SAt2	.56**	-.47**	.53**	-.55**	-.31	.86**	.26	-.09	.08	-.04	1	
ATt2	.37	-.34	.65**	-.36	.05	.63**	.08	-.16	.17	-.01	.66**	1
INt2	.61**	-.40*	.31	-.58**	-.04	.80**	.18	-.09	.09	-.07	.78**	.73**

 ISBN: 978-1-61208-229-5

As the data were normally distributed, we applied Pearson product moment correlation [25] and found several highly significant results as listed in Table 2.

Intention to further use NFC is significantly related to the participants attitude towards NFC usage at t2 ($r = .80$, $p < .01$) as well is satisfaction ($r = .78$, $p < .01$). Social influence, facilitating conditions, self-efficacy and anxiety derived from UTAUT did not show any significant influence on other constructs.

Regarding TAM constructs perceived usefulness and perceived ease of use, the results are quite diverging. Perceived ease of use has no significant effects whereas perceived usefulness has highly significant effects at both points of measurement (t1, $r = -.40$, $p < .05$; t2, $r = -.58$, $p < .01$). The constructs from expectation-disconfirmation theory did all show at least one significant correlation with other constructs. Especially the construct satisfaction is in a highly significant correlation with attitude towards NFC usage ($r = .86$, $p < .01$) and the intention to further use NFC ($r = .78$, $p < .01$). These results indicate that expectation disconfirmation theory is more appropriate in the context of NFC acceptance than the other tested instruments as it provided more significant correlations among the constructs.

Additionally, we asked the participants for reasons why task completion was difficult/easy for them, whether they consider NFC appropriate for that particular task and further comments. 434 text items were collected in the course of that and analyzed concerning their content. The two main TAM constructs, ease of use and usefulness, occurred rather often in the user comments. Usefulness was mentioned 34 times and ease of use was addressed even 53 times, which represents more than 12 % of all comments. Nevertheless other topics were more important to the participants. The ability of NFC to act as a time-saver was named in 78 comments (18 %). Another 71 comments dealt with design issues such as font size, color, haptic characteristics etc. Other often named issues were content control concerning transferred data, performance, costs, compatibility with other technologies, fun and opportunities to automat processes. Moreover, the participants provided detailed information concerning the exact form of the different acceptance criteria like what exactly means easy to use to them. Other valuable information in the user comments were product suggestions. The most prominent suggestions were NFC applications for museums, business applications for employees of a company, library applications and NFC YouTube links.

According to these results research question two is answered as following: There are similarities between acceptance factors gathered by means of traditional acceptance research instruments and those from qualitative research, but the information users provide beyond standardized questionnaires is further detailed and also more design relevant.

These results indicated that the constructs tested in traditional acceptance research are important topics but often participants are biased because of the limited number of possible answers. A participant who highly agrees with a certain statement in a standardized questionnaire will not necessarily name this item as an important factor for further usage of the technology. Research question one can be answered as following: It is possible to apply traditional acceptance research instruments, but it will cause a loss of valuable information and neglects important acceptance factors.

V. CONCLUSION AND FUTURE WORK

Generalizability of our results is of course limited due to the relatively low number of participants, which was caused by the requirements of qualitative research but nevertheless our results indicate that traditional methods of technology acceptance research only show limited ability to capture participants' opinions concerning interactivity enabling technologies such as NFC. Acceptance of NFC seems to be a very dynamic issue and therefore expectation-disconfirmation theory provided the best results due to its dynamic (two-step) data gathering process. Standardized questionnaires are extremely useful instruments for technology acceptance research as they enable collection of numerous user opinions, but at the same time they hamper detection of really valuable information, which is uncovered by means of qualitative data gathering such as interviews or thinking aloud.

The challenge for the future will be a combination of both approaches, which enables exploration of many user opinions concerning their actual thoughts not limited to a small number of possible constructs. We, therefore, believe that it will be necessary to foster methods of automated text analysis in the field of technology acceptance research as users are providing us with an incredible amount of textual information concerning their experiences with technology in form of user generated content publicly available on the internet.

First attempts to apply this kind of data gathering methods on technology acceptance problems [16] did show that automated text analysis can be a very useful instrument and also provides in-depth insights into users actual opinions concerning technologies. Our next steps, therefore, are the further development of an automated text analysis framework in the context of technology acceptance research as well as a comparative analysis of methods available in the area of technology acceptance research in order to find out which are most appropriate for interactive technologies.

REFERENCES

[1] J. Kjeldskov and J. Stage, "New techniques for usability evaluation of mobile systems," International Journal of Human-Computer Studies, vol. 60 (5-6), 2004, pp. 599-620, doi:10.1016/j.ijhcs.2003.11.001.

[2] E. Platzer, "A critical review of user acceptance research in the area of mobile services," Libri: international journal of libraries and information services, vol. 59 (4), 2009, pp. 213-227, doi: 10.1515/libr.2009.019.

[3] E. Platzer and O. Petrovic, "Approaches to address the lack of relevance in technology acceptance research," Proceedings of the 21st Central European Conference on Information and Intelligent Systems, 2010, pp. 289-296.

[4] E. Platzer, "A framework to support the design of mobile applications," Proceedings of the International Conference on Computer Networks and Mobile Computing (ICCNMC 2010), 2010, pp. 290-296.

 ISBN: 978-1-61208-229-5

[5] F.D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," MIS Quarterly, vol. 13(3), 1989, pp. 319-340.

[6] E. Platzer and O. Petrovic, "Development of technology acceptance research for mobile services," Proceedings of the 33rd International Convention on Information and Communication Technology, Electronics and Microelectronics (Mipro DE), 2010, pp. 70-76.

[7] H. Sun and Z. Ping, "A methodological analysis of user technology acceptance," Proceedings of the 37th Annual Hawaii International Conference on System Sciences, (5-8), 2004, pp. 1-10, doi:10.1109/HICSS.2004.1265621

[8] H. Sun and Z. Ping, "Applying Markus and Robey's Causal Structure to Examine User Technology Acceptance Research: A New Approach," Journal of Information Technology Theory and Application, vol. 8 (2), 2006, pp. 21-40.

[9] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward a Unified View," MIS Quarterly, vol. 27 (3), 2003, pp. 425-478.

[10] R.L. Oliver, "A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions," Journal of Marketing Research (JMR), vol. 17 (4), 1980, pp. 460-469.

[11] A. Bhattacherjee and G. Premkumar, "Understanding Changes in Belief and Attitude toward Information Technology Usage: A Theoretical Model and Longitudinal Test," MIS Quarterly, vol. 28 (2), 2004, pp. 229-254.

[12] M.-C.Hung, H.-G. Hwang, and T.-C. Hsieh, "An exploratory study on the continuance of mobile commerce: an extended expectation-confirmation model of information system use," International Journal of Mobile Communications, vol. 5 (4), 2007, pp. 409-422.

[13] R. Fensli, P. E. Pedersen, T. Gundersen, and O. Hejlesen, "Sensor Acceptance Model – Measuring Patient Acceptance of Wearable Sensors," Methods of Information in Medicine, vol. 47 (1), 2008, pp. 89-95, doi:10.3414/ME9106.

[14] J. Blom, J. Chipchase, and J. Lehikoinen, "Contextual and cultural challenges for user mobility research," Commun. ACM, vol. 48 (7), 2005, pp. 37-41.

[15] C. Page, "Mobile research strategies for a global market," Commun. ACM, vol. 48 (7), 2005, pp. 42-48.

[16] E. Platzer and O. Petrovic, "Learning Mobile App Design From User Review Analysis," International Journal of Interactive Mobile Technologies, vol. 5 (3), 2011, pp. 43-50.

[17] R. Jain, A. Shivaprasad, D. Lelescu, and X. He, "Towards a model of user mobility and registration patterns," SIGMOBILE Mob. Comput. Commun. Rev., vol. 8 (4), 2004, pp. 59-62.

[18] G. Resta and P. Santi, "WiQoSM: An Integrated QoS-Aware Mobility and User Behavior Model for Wireless Data Networks," IEEE Transactions on Mobile Computing, vol. 7 (2), 2008, pp. 187-198, doi:10.1109/TMC.2007.70728.

[19] M. McNett and G.M. Voelker, "Access and mobility of wireless PDA users," SIGMOBILE Mob. Comput. Commun. Rev., vol. 9 (2), 2005, pp. 40-55.

[20] I. Junglas, "On the usefulness and ease of use of location-based services: insights into the information system innovator's dilemma," Int. J. Mob. Commun., vol. 5 (4), 2007, pp. 389-408.

[21] F.D. Davis and V. Venkatesh, "Toward preprototype user acceptance testing of new information systems: implications for software project management," IEEE Transactions on Engineering Management, vol. 51 (1), 2004, pp. 31-46, doi:10.1109/TEM.2003.822468.

[22] M. Isomursu, M. Tähti, S. Väinämö, and K. Kuutti, "Experimental evaluation of five methods for collecting emotions in field settings with mobile applications," International Journal of Human-Computer Studies, vol. 65 (4), 2007, pp. 404-418, doi:10.1016/j.ijhcs.2006.11.007.

[23] A. Kaikkonen, A. Kekäläinen, M. Cankar, T. Kallio, and A. Kankainen, "Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing," Journal of Usability Studies, vol. 1 (1), 2005, pp. 4-16.

[24] A. Bhattacherjee, "Understanding Information Systems Continuance: An Expectation-Confirmation Model," MIS Quarterly, vol. 25 (3), 2001, pp. 351-370.

[25] J. L. Rodgers and W. A. Nicewander. "Thirteen ways to look at the correlation coefficient," The American Statistician, vol. 42 (1), 1988, pp. 59–66.

 ISBN: 978-1-61208-229-5

Channel-Matched Space-Time Code Selection and Adaptive Modulation for Rayleigh Fading Channels

Said El-Khamy
Department of Electrical Engineering
Faculty of Engineering, Alexandria University
Alexandria, Egypt
elkhamy@ieee.org

Amr El-Helw, Sara Zaalik
Electronics and Communication Department
Arab Academy for Science and Technology (AAST)
Alexandria, Egypt
A.M.El-Helw@aast.edu, sarazaalik@gmail.com

***Abstract*— Space-Time Block Codes represent an exciting development in the field of wireless communication. It is a promising technique to increase the data rates with minimal decoding complexity. In this paper, we present an adaptive transmission system for Rayleigh fading Multiple Input Multiple Output channels. Code selection and adaptive modulation are combined together by two feedback scenarios. Channel State Information is sent back from the receiver to the transmitter controlling three feedback decision bits. By applying these adaptive techniques, the system adjusted its performance with the channel conditions. As a result, signal transmission with an improved coding gain and diversity near to the maximum diversity order is achieved. The proposed system has shown enhanced bit error rate performance and better outage data throughput, when compared with non adaptive systems.**

Keywords--Space-Time Block Codes; Multiple Input Multiple Output; Channel State Information; feedback; Adaptive Modulation.

I. INTRODUCTION

Third-generation wireless communication systems are required to provide data bit rates of up to 2 Mbits/s. Recently released fourth-generation standards will push these rates even higher, possibly over 100 Mbits/s [1]. To support such high rates, multiple antennas can be employed to increase channel capacity.

As a result of Space Time Coding (STC) [4], the transmitted signal resists the multipath fading effects that increases the Bit Error Rate (BER) by depending on both time and space diversity. It was shown in [4] that an orthogonal full-rate design, offering full diversity for any arbitrary complex symbol constellation, is limited to the case of two transmit antennas. Data-rate or decoding simplicity must be sacrificed if the number of transmit antennas is increased.

Most Space Time Block Codes (STBCs) are designed under the assumption that the transmitter has no knowledge about the channel. On the other hand, it has been shown in [3] that an outage performance with perfect Channel State Information (CSI) available at the transmitter and at the receiver is better compared to the case when only the receiver has perfect knowledge of the channel.

When applying adaptive modulation, great enhancements in the system performance are achieved. Constantly changing the modulation scheme with the varying channel conditions and thus, yielding higher data throughput when compared with the non adaptive systems [7]. According to [4], an orthogonal complex 4x4 code matrix is not full rate on the other hand, the Extended Alamouti 4x4 near orthogonal code matrix is full rate

Thus sacrificing the code orthogonality, code selection technique is used to overcome this problem. The proposed system combines adaptive modulation and code selection techniques depending on the CSI available at both the transmitter and receiver, this combined system outstands the non-adaptive systems in both bit error rate (BER) results and outage data throughput.

The paper is organized as follows. Section II reviews the simple 2x1 Alamouti scheme [4] and the Extended Alamouti scheme with code selection technique [3]. Section III presents the proposed system combining adaptive modulation with code selection techniques. The system simulation and results will be presented in Section IV. Finally, Section V contains the paper conclusion.

II. ALAMOUTI AND EXTENDED ALAMOUTI SCHEMES

A. Alamouti Scheme

In 1998, the preliminary form of STBC was introduced by Alamouti [4]. It linearly and orthogonally encodes a data stream and transmits it simultaneously across the channel. The encoder takes a block of two modulated symbols S_1 and S_2 in each encoding operation and maps them to two transmit antennas according to a code matrix given in (1). After transmission, the data stream is successfully extracted at the receiver due to the orthogonal encoding.

$$S = \begin{pmatrix} S_1 & S_2 \\ S_2^* & -S_1^* \end{pmatrix} \quad (1)$$

At the receiver, the signals are expressed as:

$$R_1 = h_1 S_1 + h_2 S_2 + n_1 \quad (2)$$
$$R_2 = h_1 S_2^* - h_2 S_1^* + n_2 \quad (3)$$

Or, written in the vector form, as mentioned in [5]

$$r = Sh+n \quad (4)$$

where ***n*** represents Gaussian noise.

B. 4x1 Extended Alamouti Scheme and code selection

The Extended Alamouti Space Time Block Coding (EA-STBC) uses four transmit antennas and one receiver antenna [3]; as a result, four symbols are transmitted each time slot. The code matrix is generated as a result of "alamoutisation" of the basic Alamouti code mentioned in (1) [2]. In other words, two Alamouti codes are used to

 ISBN: 978-1-61208-229-5

build up the EA-STBC for four transmit antennas. The resulting code extends over four time slots and is described in [3][5] by the following signal matrix,

$$S_1 = \begin{pmatrix} S_1 & S_2 & S_3 & S_4 \\ S_2^* & -S_1^* & S_4^* & -S_3^* \\ S_3^* & S_4^* & -S_1^* & -S_2^* \\ S_4 & -S_3 & -S_2 & S_1 \end{pmatrix} . \quad (5)$$

The received signals within four successive time slots, assuming one receiver antenna, are given in [3] as**,**

$$r_1 = S_1 h_1 + S_2 h_2 + S_3 h_3 + S_4 h_4 + n_1 \quad (6)$$
$$r_2 = S_2^* h_1 - S_1^* h_2 + S_4^* h_3 - S_3^* h_4 + n_2 \quad (7)$$
$$r_3 = S_3^* h_1 + S_4^* h_2 - S_1^* h_3 - S_2^* h_4 + n_3 \quad (8)$$
$$r_4 = S_4 h_1 - S_3 h_2 - S_2 h_3 + S_1 h_4 + n_4. \quad (9)$$

With complex conjugation of (7) and (8), we obtain,

$$y_2 = r_2^* , \quad n_2 = n_2^* \quad (10)$$
$$y_3 = r_3^* , \quad n_3 = n_3^* \quad (11)$$

Resulting in (12), the matrix equation representing the transmission scheme,

$$y = H_{v1} s + n \quad (12)$$

where H_{v1} is the virtual effective channel matrix and is equal to,

$$\boldsymbol{H_{v1}} = \begin{pmatrix} h_1 & h_2 & h_3 & h_4 \\ -h_2^* & h_1^* & -h_4^* & h_3^* \\ -h_3^* & -h_4^* & h_1^* & h_2^* \\ h_4 & -h_3 & -h_2 & h_1 \end{pmatrix} \quad (13)$$

It is shown in [3] that H_{v1} is nearly orthogonal. As mentioned in [4] that fully orthogonal codes achieve full diversity gain. So, to overcome the problem of near orthogonal codes, the code selection technique is proposed in [3].

The signal transmission is described by the vector form discussed in (4) where **r** is the (4x1) vector of the received signals. **S** is the space time block code that could be either S_1 as defined in (5) or S_2 defined in (14) depending on a feedback bit b_3, and finally, **n** is the (4x1) noise vector [5]. In [3][5], the EA-STBCs are generated from each others by changing the signs, which is obvious in the generated code named S_2 in (14).

$$S_2 = \begin{pmatrix} -S_1 & S_2 & S_3 & S_4 \\ -S_2^* & -S_1^* & S_4^* & -S_3^* \\ -S_3^* & S_4^* & -S_1^* & -S_2^* \\ -S_4 & -S_3 & -S_2 & S_1 \end{pmatrix} \quad (14)$$

Similarly, H_{V2} is generated if S_2 is used,

$$\boldsymbol{H_{v2}} = \begin{pmatrix} -h_1 & h_2 & h_3 & h_4 \\ -h_2^* & -h_1^* & -h_4^* & -h_3^* \\ -h_3^* & -h_4^* & -h_1^* & h_2^* \\ h_4 & -h_3 & -h_2 & -h_1 \end{pmatrix} \quad (15)$$

As in [5], we obtain

$$G = H_{vi}^H H_{vi} = H_{vi} H_{vi}^H$$

$$= h^2 \begin{pmatrix} 1 & 0 & 0 & \mathbf{X_i} \\ 0 & 1 & -\mathbf{X_i} & 0 \\ 0 & -\mathbf{X_i} & 1 & 0 \\ \mathbf{X_i} & 0 & 0 & 1 \end{pmatrix} \quad (16)$$

where i =1 or 2 and $h^2 = h_1^2 + h_2^2 + h_3^2 + h_4^2$ (17)

$$X_1 = \frac{2\,\mathrm{Re}\left(h_1 h_4^* - h_2 h_3^*\right)}{h^2}, \quad \text{when } S_1 \text{ is sent} \quad (18)$$

And

$$X_2 = \frac{2\,\mathrm{Re}\left(-h_1 h_4^* - h_2 h_3^*\right)}{h^2}, \quad \text{when } S_2 \text{ is sent} \quad (19)$$

It is known from [6] that **G** should be an identity matrix to achieve full diversity and an optimum system performance. But, on the other hand, if G is not an identity matrix as in (16), channel dependent parameter named X_i appears leading to interference between the four channel parameters. Thus, X_i should be as small as possible to reach near orthogonality of STBC used**.**

Thus, code selection is applied between the two transmission codes S_1 and S_2. By computing the values of X_1 and X_2 from (18) and (19), respectively, the system returns feedback bit $\boldsymbol{b_3}$ to the transmitter to select the code block corresponding to the minimum value of *X*.

 ISBN: 978-1-61208-229-5

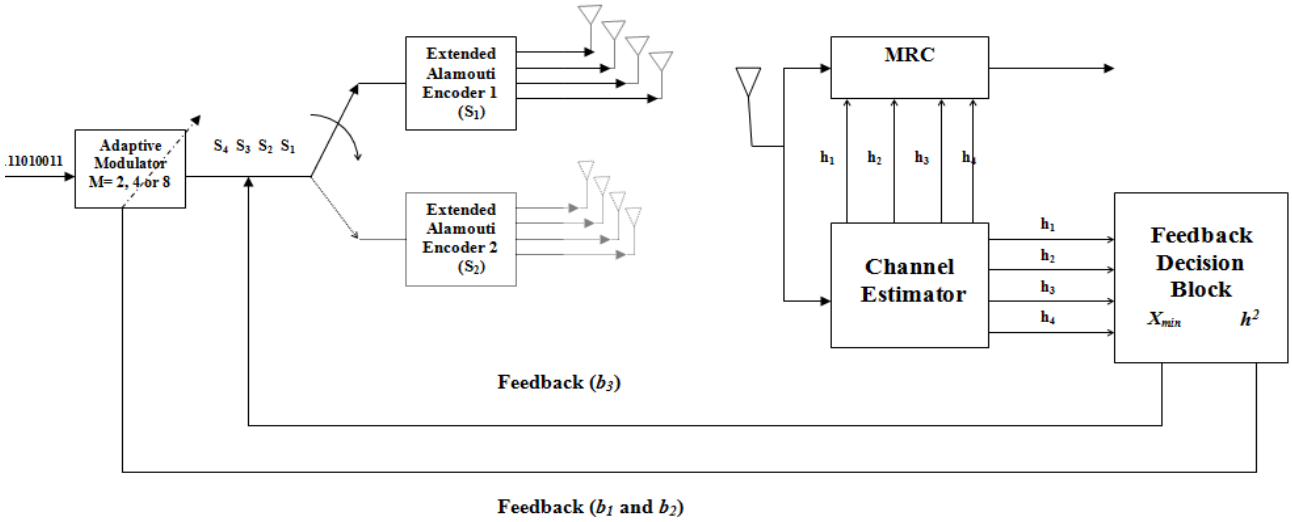


Figure 1. Combined Adaptive Modulation and Code Selection Techniques

III. Proposed Combined Adaptive Modulation and Code Selection Techniques

The combined adaptive transmission system proposed in this paper requires full knowledge of CSI at both the transmitter and the receiver. Figure 1 presents the block diagram of the proposed combined system. Binary data is fed into an adaptive M-ary modulator set to three modulation states, BPSK, QPSK or 8-ary PSK. Switching between these modulation techniques requires two feedback bits $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$. The data is then encoded with either one of two Extended Alamouti block codes S_1 or S_2. Switching between these block codes requires a feedback bit $\boldsymbol{b_3}$. The encoded data is then transmitted through four time slots, using four different antennas. At the receiver end, the transmitted signals are received using one receiver antenna. Finally, Maximum Ratio Combining (MRC) is used for signal detection at the receiver.

Choosing between S_1 and S_2 at the transmitter is considered a type of system adaptation, or, in other words, matching the code selection process with the value of channel dependent interference parameter X_i given in (18) and (19). Adaptive modulation is also applied which denotes the matching of the modulation technique with the channel conditions. Higher order modulation is assigned to the system at poor channel conditions. In this way we could benefit from the channel response, whereas the channel is considered as gain for the transmitted signal. In this paper, we have investigated three modulation techniques, BPSK, QPSK and 8-PSK, therefore requiring two feedback bits $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ implemented between the receiver end and the modulator in the transmitter as shown in Figure 1.

For acquiring the CSI, two thresholds are defined to assign a modulation scheme to the system using bits $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$; the third feedback bit $\boldsymbol{b_3}$ returns information on the code block that will be used to encode the modulated data.

From (17), $\boldsymbol{h^2}$ is easily computed, and, according to its value, the suitable modulation is chosen. As a result, the adaptive modulation system controls the outage data rate. In the following section, performance of the proposed system which combines adaptive modulation and code selection will be investigated.

IV. Simulation and Results

In our simulations, we have used flat Rayleigh fading channel remaining constant during the transmission of each code block [3]. At the receiver side, we have used MRC receiver. The BER results have been averaged over 10^5 realizations of i.i.d channel matrix. We simulated MIMO systems with two and four transmit antennas and a single receive antenna. BPSK, QPSK [3] and 8-PSK modulation techniques are used in the adaptive modulation process. In the simulations of the adaptive modulation systems, the BER and the modulation order (M) are averaged for each value of Signal to Noise Ratio (SNR). The results are introduced in three parts; part A presents the processes of adaptive modulation and code selection while, the BER results are presented in part B. To visualize the data throughput more clearly and to support the enhanced BER results, Figure of Merit is calculated, which is presented in part C.

A. Adaptive modulation and code selection Processes

This section presents the behavior of the proposed combined system for a given sample of ten time frames for the same SNR.

In Figure 2, the normalized channel response is represented against time frame. It is quiet obvious that the channel is varying at each frame. By extensive simulation of the proposed system according to the simulation environment mentioned at the beginning of this section, it was noticed that the values of $\boldsymbol{h^2}$ lies between a certain range of numbers. Two middle values were chosen as the threshold values to control the adaptive modulation process. During the time frames that contain values of the normalized channel beneath threshold 1, BPSK (M=2) is assigned. While QPSK (M=4) is used during the time frames containing values of the normalized channel in between threshold 1 and threshold 2. Finally, 8-PSK (M=8), is assigned otherwise (above threshold 2) as shown below.

 ISBN: 978-1-61208-229-5

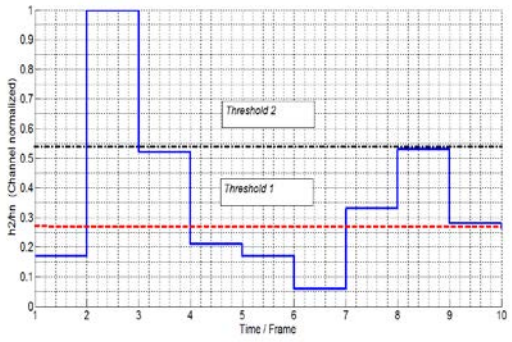


Figure 2. Channel Response Normalized (h^2/h_n) Against Time/Frame

On the other hand, Figure 3 shows the system response (from the adaptive modulation aspect) towards the variation of the normalized channel values- presented in Figure 2 for the exact ten time frames leading to variation in the modulation order used at each frame.

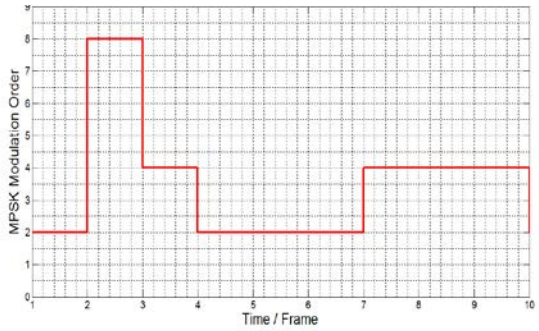


Figure 3. MPSK Modulation Order at each Time/Frame

Figure 4 represents the values of the channel dependent parameter, Xi, against the same time frames presented in Figures 2 and 3. With varying values of X_1 and X_2, code switching between S_1 and S_2 is applied by the system as shown below. For easy visualization of this process, S_1 is represented by 1 and S_2 is represented by 1.5.

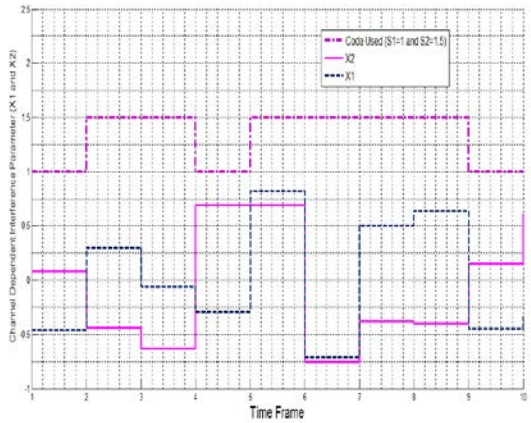


Figure 4. Code Selection Technique

B. BER Results

For the system setup mentioned in the beginning of Section IV, this section shows the resulting performances (BER against SNR) for the proposed combined system against non adaptive systems. Figure 5 shows the resulting BER against SNR for the 2x1 simple Alamouti scheme that uses adaptive modulation against non-adaptive system (no code selection is applied in both systems). Performance improvement is obvious, for the same BER the adaptive system is almost 7dB better than the non-adaptive one.

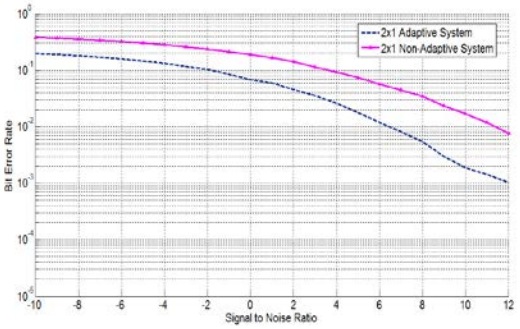


Figure 5. BER for 2x1 Adaptive Alamouti scheme against 2x1 Non-Adaptive Alamouti scheme

Figure 6 shows the BER results for the proposed system (adaptive modulation and code selection techniques) against the non-adaptive system that uses only QPSK modulation. The proposed system gives 7dB SNR improvement than the non adaptive system.

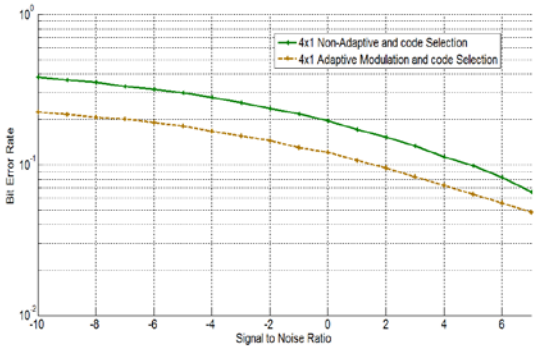


Figure 6. BER for 4x1 Adaptive system against 4x1 non adaptive system.

Figure 7 shows the BER performance improvements for the proposed combined adaptive system that uses three modulation schemes (at the same time) against non adaptive systems using BPSK, QPSK and 8-PSK each at a time, The results shows at least 4 dB of SNR improvement as shown in the figure.

 ISBN: 978-1-61208-229-5

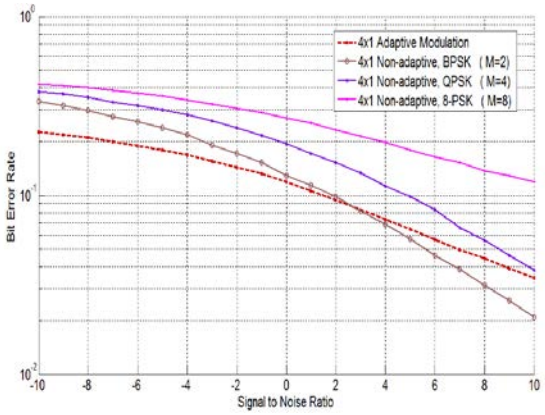


Figure 7. The 4x1 combined system against non adaptive systems using BPSK, QPSK and 8-PSK

At almost 3 dB, the performance of the 4x1 BPSK outstands the proposed combined system which is expected as the BPSK should have the optimum performance (lower BER) for all ranges of SNR on the account of the data throughput. On the other hand, the proposed combined system shows better throughput than the other non-adaptive systems as will be shown in Figure 8.

C. *Figure of Merit*

Figure of merit is an indication of the data throughput of the system. It is drawn against the SNR and it is calculated by the following formula,

$$\text{Figure of Merit} = M^* (1\text{-BER}) \qquad (20)$$

where M is the modulation order.

Figure 8 shows the resulting figure of merit for 4x1 proposed combined adaptive system against non-adaptive BPSK and QPSK. As clearly shown, the adaptive system shows better figure of merit than the other systems therefore better throughput. For the adaptive system, in the calculation of the above formula, M is averaged along the whole adaptive modulation process (as mentioned earlier). Its quiet clear that the figure of merit tends to M in each case where M_{max}=M.

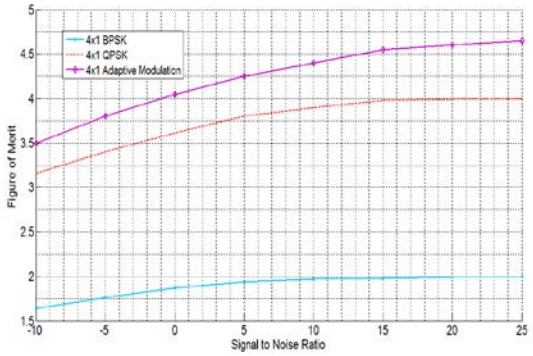


Figure 8. Figure of Merit results for the 4x1 Adaptive System Against 4x1 non Adaptive Systems.

V CONCLUSION

By combining both adaptive modulation and code selection to the 4x1 Extended Alamouti system, performance improvements are visible in all the presented results. With assigning higher order modulation to the system when $\boldsymbol{h}^2$ is above threshold 2, it is considered as a gain for the transmitted signal therefore the signal is more robust to noise. Figure of Merit results indicate better throughput for the adaptive system by calculating the average modulation order (Maverage) through out the process which will be greater than that of the BPSK and QPSK individually. In addition, code selection between the non- orthogonal codes S_1 and S_2 to reach the code with minimum value of X_i, in other words, the near orthogonal code. Changing the receiver and the number of the receiver antennas could be considered later on to achieve near optimum performances.

ACKNOWLEDGMENT

I would like to express my deep gratitude to the reviewers as well as all my professors and colleagues.

REFERENCES

[1] N. S. J. Pau, “Robust high throughput Space-Time Block Coded MIMO systems”, Ph.D. Thesis, University of Canterbury, Christchurch, New Zealand, June 2007.

[2] J. F. Torras, “New Hybrid Automatic Repeat Request (HARQ) scheme for a 4x4 MIMO system, based on The Extended Alamouti Quasi-orthogonal Space-Time Block Coding (Q-STBC), in invariant and variant fading channel”, M.Sc. Thesis ,New Jersey Institute of Technology, May 2006.

[3] B. Badic, M. Rupp, and H. Weinrichter, “Adaptive channel-matched Extended Alamouti Space-Time Code exploiting partial feedback”, ETRI Journal, Vol. 26, No. 5, pp. 443-451 , October 2004

[4] V. Tarokh, H. Jafarkhani, and A.R. Calderbank, “Space-Time Block Codes from orthogonal design ” IEEE Trans. Inf. Theory, vol. 45, pp. 1456-1467, July 1999.

[5] M. Rupp and C. F. Mecklenbräuker, “On Extended Alamouti schemes for Space-Time Coding”, The 5th International Symposium on Wireless Personal Multimedia Communications, Honolulu, pp. 115-119, October 2002.

[6] C. F. Mecklenbräuker and M. Rupp, “Flexible Space-Time Block Codes for trading quality of service against data rate in MIMO UMTS”, EURASIP J. Applied Signal Processing, Special Issue on MIMO Communication and Signal Processing, vol. no. 5, pp. 662-675, May 2004.

[7] J. Yang, N. Tin, and A. K. Khandani, "Adaptive modulation and coding in 3G wireless systems", in proceedings of the 56th IEEE Vehicular Technology conference, vol.1, pp. 24-28, September 2002.

 ISBN: 978-1-61208-229-5

Implications of using a Large Initial Congestion Window to Improve mSCTP Handover Delay

Johan Eklund
Department of Computer Science
Karlstad University
Sweden
Email: johan.eklund@kau.se

Karl-Johan Grinnemo
Department of Computer Science
Karlstad University
Sweden
Email: karl-johan.grinnemo@kau.se

Anna Brunstrom
Department of Computer Science
Karlstad University
Sweden
Email: anna.brunstrom@kau.se

***Abstract*—The currently rather heterogeneous wireless landscape makes handover between different network technologies, so-called vertical handover, a key to a continued success for wireless Internet access. Recently, an extension to the Stream Control Transmission Protocol (SCTP) – the Dynamic Address Reconfiguration (DAR) extension – was standardized by IETF. This extension enables the use of SCTP for vertical handover. Still, the way vertical handover works in SCTP with DAR makes it less suitable for real-time traffic. Particularly, it takes a significant amount of time for the traffic to ramp up to full speed on the handover target path. In this paper, we study the implications of an increased initial congestion window for real-time traffic on the handover target path when competing traffic is present. The results clearly show that an increased initial congestion window could significantly reduce the transfer delay for real-time traffic, provided the fair share of the available capacity on the handover target path is sufficiently higher than the send rate required by the real-time flow. Additionally, we notice that this performance gain comes without penalizing the competing traffic.**

***Keywords*-SCTP; dynamic address reconfiguration; video; mobility; handover; congestion control; slow start**

I. INTRODUCTION

Wireless networks comprises a variety of technologies, e.g., cellular (3G/4G), WiFi, and WiMAX. Additionally, the number of terminals with multiple wireless interfaces to access Internet is rapidly increasing. Ubiquitous connectivity is made possible by the ability for a single device to roam between heterogeneous networks. The goal for this, so-called, vertical handover is to be transparent to the end user.

The Stream Control Transmission Protocol (SCTP) [1], with its multihoming feature, and its extension for Dynamic Address Reconfiguration (DAR) [2], a.k.a. mSCTP, has become a promising alternative for vertical handover. The DAR mechanism enables for SCTP to dynamically add and remove IP addresses to an ongoing session.

However, to be able to provide a seamless handover, mSCTP needs to have an efficient handover detection mechanism, particularly a mechanism that anticipates the loss of connectivity to the current access point. Furthermore, mSCTP has to start up swiftly on the handover target path. Although a vertical handover detection mechanism is indeed important, several works have already studied this issue. Thus, in this paper it is assumed that we have an optimized handover mechanism that contributes marginally to the handover delay. Instead, this paper focuses on the second issue: the startup on the handover target path. In fact, earlier studies [3], [4] have shown that even in a scenario with an ideal handover detection mechanism, the mobile terminal may experience a non-negligible service disruption, due to the startup phase on the handover target path. We have also seen that a way to mitigate this startup delay after handover could be to increase the initial congestion window (*init_cwnd*) on the handover target path.

This paper is a continuation of our previous study on using an increased *init_cwnd* to improve the vertical handover performance of mSCTP. Particularly, the paper studies the implications of an increased *init_cwnd*, in terms of latency and fairness, in a situation with competing traffic. We study the impact of using an *init_cwnd* for the mSCTP traffic, on both mSCTP and the competing traffic.

The paper considers the effects on real-time video traffic (H.264, HQ) in scenarios, which are intended to model a vertical handover from an arbitrary wireless network to a cellular 3G or 4G network. The competing traffic on the handover target path consists of elastic background traffic. In this case, one or several TCP flows.

The results of our study suggest that an increased *init_cwnd* could significantly reduce the handover delay, provided the capacity required by the video traffic is significantly lower than its fair share on the handover target path. The results also suggest that this performance gain has no effect on the short-time fairness to competing traffic.

The remainder of this paper is organized as follows. Section II gives an overview of SCTP with a focus on its support for multihoming. Next, Section III discusses the setup and methodology used in our experimental study. The results from the study are presented and discussed in Section IV. Section V surveys related work. Finally, Section VI concludes the paper and briefly mentions ongoing and future work.

II. PRELIMINARIES

The Stream Control Transmission Protocol (SCTP), originally developed as a transport protocol to serve telephony call-control signaling, is today standardized as a general-purpose transport protocol in RFC 4960 [1]. Still, new developments on SCTP are being made. Several of the current standardization

 ISBN: 978-1-61208-229-5

activities are described in Dreibholz et al. [5]. Some extensions have been standardized in separate RFCs [2], [6], [7].

SCTP does in many ways mimic TCP SACK [8]; It is a reliable, connection-oriented transport protocol that offers a selectively-acknowledged, non-duplicated transfer of packets. Further, it uses window-based congestion- and flow-control mechanisms that essentially work the same as in TCP SACK. In default mode, data is delivered to the application in ordered mode, while unordered or partial-ordered delivery, which could be suitable for real-time traffic, is optional. To distinguish an SCTP connection from a connection in TCP, SCTP uses the term "association".

However, one major extension to SCTP is the multihoming feature, which implies that an association is able to connect to several IP addresses at both the source and destination endpoints. Normally, SCTP selects one of its peer's destination addresses as the primary destination address. The remaining addresses serve as alternate or backup addresses. If the primary destination address becomes unavailable, a failover procedure takes place, which results in the traffic being re-routed to one of the alternate addresses.

The Dynamic Address Reconfiguration (DAR)) [2] extension of SCTP enables for an SCTP endpoint to dynamically alter its IP addresses during the lifetime of an association. Apart from permitting IP addresses to be dynamically added to and removed from associations, the DAR extension provides for an SCTP endpoint to explicitly change its peer's primary destination address. The DAR extension is key to facilitate transport-level mobility and several transport-level mobility solutions that build on SCTP extended with DAR (mSCTP) have been proposed [9], [10], [11], [12].

III. Experimental Methodology

The experiment models a scenario where a mSCTP-based, video session is handed over to a network, where elastic TCP traffic is present. Our scenario is illustrated in Figure 1. Initially, one or several video sessions are set up between a stationary and a mobile terminal, utilizing Path 1.

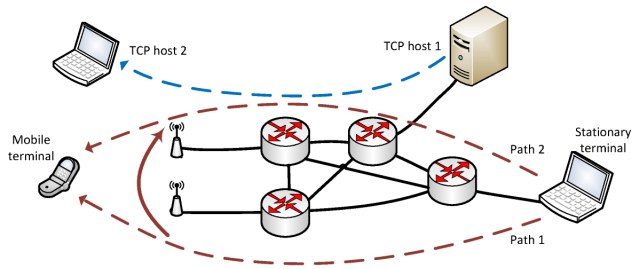


Fig. 1. Schematic view of our experimental scenario.

Some time after the video sessions has left the slow-start phase, the sessions are handed over to Path 2, where the bottleneck section of the path is shared between the mSCTP video sessions and the competing traffic.

The experiment comprised end hosts, which ran real implementations of mSCTP and TCP. The network was emulated utilizing the KauNet network emulator [13], which is an extension to the Dummynet emulator [14]. A view of the experiment setup is depicted in Figure 2.

The video traffic mimicked video traffic from a standard definition (H.264, HQ) video clip, an episode of the "Horizon Talk Show", made available by the Arizona State University Video Trace Library [15], [16]. The mSCTP traffic was generated by a Custom-made Traffic Generator (CTG). The clocks of the end hosts were synchronized, and all messages were time stamped at departure from the sending application and at reception by the target application.

The competing traffic on the handover target path consisted of bulk TCP flows. The motivation behind this choice was to get a view of the impact of a large *init_cwnd* on the competing traffic, as well as to study the behavior of a video flow in a scenario with competing traffic. The competing traffic was generated by the Iperf [17] traffic generator.

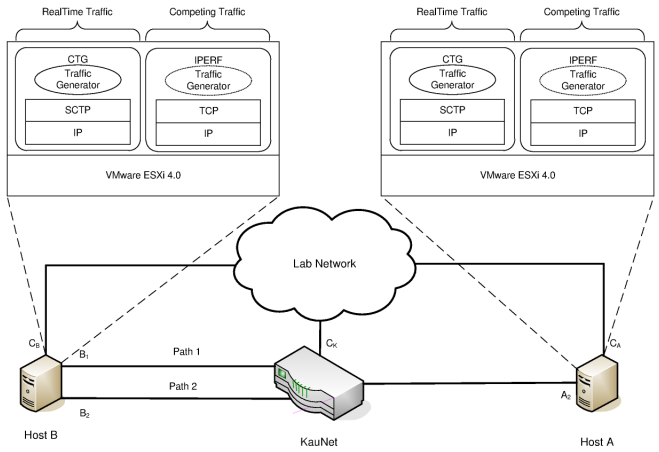


Fig. 2. Experiment setup.

A. Parameters

The video traffic was sent at Constant Bit Rate (CBR)[1] and the average send rate was 1.53 MBps, generated at a rate of 6395 bytes messages every 33 ms. The networks considered in this study were a high latency, low capacity (*lc*) network (RTT 300 ms, bandwidth 6 Mbps) with characteristics similar to a 3G cellular network, and a lower latency and "high" capacity (*hc*) network (RTT 40 ms, bandwidth 45 Mbps), i.e., characteristics more like a 4G cellular network.

The most important parameters of our experiments are listed in Table I. Since the characteristics of the considered networks were very diverse, the number of video flows as well as the number of competing TCP flows varied between the *lc* and the *hc* experiments. Particularly, we conducted the *lc* network experiments with 1 and 2 video flows and with 1 and 2

[1]Video traffic is generally generated at Variable Bit Rate (VBR), but since we consider average results of 40 repetitions, see below, we consider CBR traffic to be appropriate.

 ISBN: 978-1-61208-229-5

competing flows, and the *hc* network experiments with 1,2 and 5 video flows and with 2, 5 and 10 competing flows.

TABLE I
PARAMETERS FOR THE DIFFERENT NETWORK TYPES

	lc	*hc*
Bandwidth (MBps)	6	45
RTT (ms)	300	40
Competing flows	1, 2	2, 5, 10
mSCTP associations	1, 2	1, 2, 5
init_cwnd (MSS)	3 (default), 10, 20, NR	3 (default), 10, 20, NR

The Maximum Segment Size (MSS) for a datagram was in the experiment set to 1500 Bytes. The experiments were run with an *init_cwnd* of 3 MSS (default), 10 and 20 MSS for the video traffic. Furthermore, to obtain an appreciation of the startup performance in a scenario where the *init_cwnd* did not impose any restriction on the video transfer, we utilized an *init_cwnd* of 50 KBytes (NR), that is an *init_cwnd* larger than the maximum number of outstanding packets in any of the considered experiments. The competing traffic was sent with the default *init_cwnd* of 3 MSS. Well aware of the ongoing discussion on the size of the router buffers [18], we conducted the experiments with a router buffer of one BDP (225 KB)[2], a typical recommendation in the literature. The send and receive buffers, as well as the maximum $ssthresh$ of the end hosts were set to large sizes, not to impose any restrictions on the video transfer.

IV. RESULTS

In every experiment, the transfer times of the individual messages in the video flow (MTTs) were measured. The MTT was measured as the time from the generation of a message by the source application until the message was received by the destination application. Additionally, in all experiments we extracted the Maximum MTT ($MMTT$). Previous studies [4] suggest that the $MMTT$ is related to the number of messages being affected by the handover delay. To be able to analyze whether or not the size of the *init_cwnd* had any impact on the competing traffic, we monitored the total throughput of the competing TCP flows. The throughput was sampled every 5 seconds. Henceforth, we call every sample a measurement point.

For normality assumptions to apply, we chose to repeat each experiment 40 times, and from the results we calculated the average value together with a 95% confidence interval. In the remainder of this section, the results from the *lc* and *hc* handover experiments are presented.

A. Handover to a low capacity network

The long transfer times, and the restricted bandwidth are parameters that are crucial for the handover performance in *lc* networks. Figure 3 represents a scenario where one video flow is handed over to the target path. The graphs represent the average *MTT*s for the messages in the video flow during startup with one competing TCP flow present. The different graphs represent the different *init_cwnd*s. In the figure, it is seen that an increase of the *init_cwnd* resulted in a reduction of the average *MTT* for a message. Further, it follows that an increased *init_cwnd* decreased the $MMTT$. Particularly, there was a major reduction in $MMTT$ when the *init_cwnd* was increased from its default value of 3 MSS up to 10 MSS, while there was a smaller, but still significant, reduction in $MMTT$ as the *init_cwnd* was further increased up to 20 MSS. The large confidence intervals were due to retransmissions of some lost messages during the startup phase. It should be noted that these results are in line with the results from our previous work without competing traffic [4].

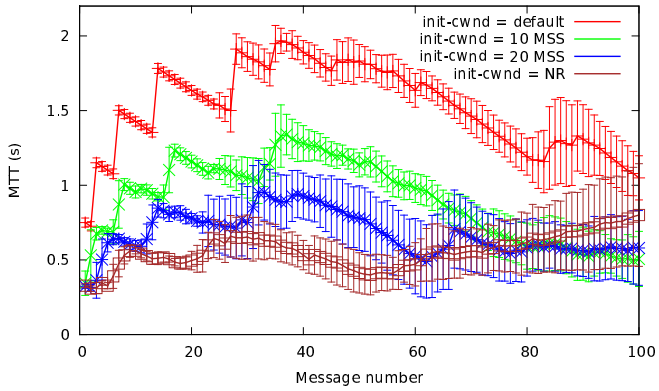


Fig. 3. MTT for one video flow. Handover of one video flow to a low capacity network with one competing flow.

Figure 4 shows the throughput evolution for the competing TCP flow as the video flow is handed over. The video flow was handed over to the target path some time after the TCP traffic was started, giving enough time for the TCP flow to reach its stationary phase.

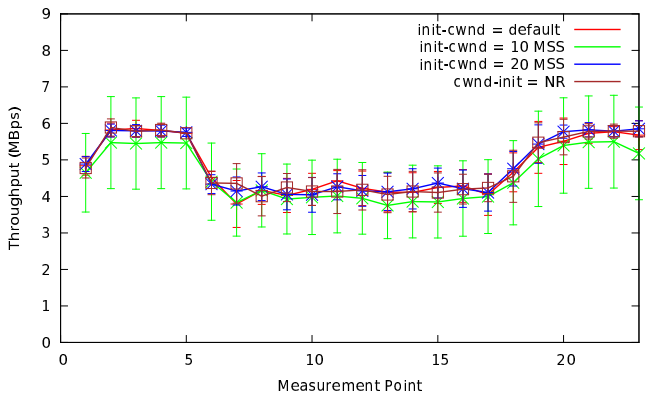


Fig. 4. Throughput for competing traffic. Handover to a low capacity network with one competing flow.

As seen from the figure, the video flow started up on the target path roughly at measurement point 4, and ended at

[2]The BDP for the *lc* and the *hc* networks turned out to be the same, although the bandwidths and the latencies differed.

 ISBN: 978-1-61208-229-5

measurement point 17. It also follows that the TCP traffic backed off appropriately to let the arriving video flow have its fair share of the bandwidth. Moreover, the figure indicates that the more aggressive startup, which was a result of an increased *init_cwnd*, had no significant impact on the performance of the competing traffic.

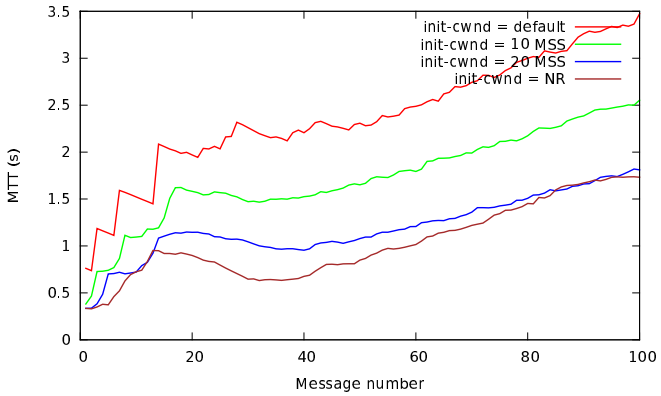


Fig. 5. MTT for one video flow. Handover of two video flows to a low capacity network with one competing flow.

Next, let us consider the experiment with two mSCTP flows being simultaneously handed over to a target path with one competing TCP flow. In this scenario, the available fair share of the bandwidth should theoretically be enough to satisfy the mSCTP flows, but, as seen in Figure 5[3], this was not the case. Instead, the MTTs increased linearly. The reason to this unexpected result was that the TCP flow did not back off appropriately, something which is seen in Figure 6.

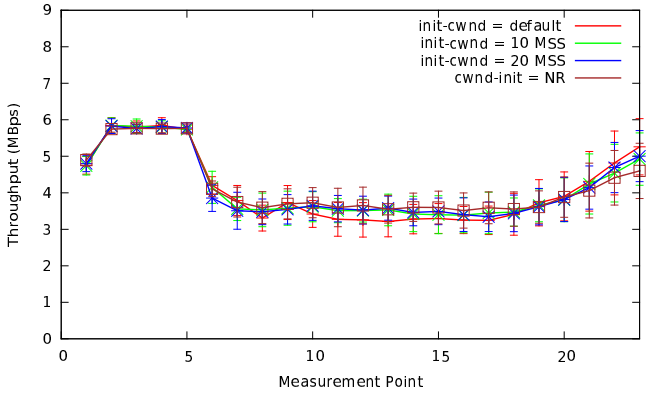


Fig. 6. Throughput for competing traffic. Handover of two video flows to a low capacity network with one competing flow.

Although the TCP flow backed off, it did not back off to more than 3.5 Mbps, something which hindered the mSCTP flows to obtain their fair shares. Still, it could be observed that at least initially, an increased *init_cwnd* resulted in a significant decrease of the experienced MTTs.

[3]For visibility purpose the confidence intervals for this experiment has been omitted. The results for the second video flow looks similar to the results for the shown flow. Thus, the second flow is not shown.

B. Handover to a high capacity network

A scenario where one video flow was handed over to a *hc* network where two competing flows were present is found in Figure 7.

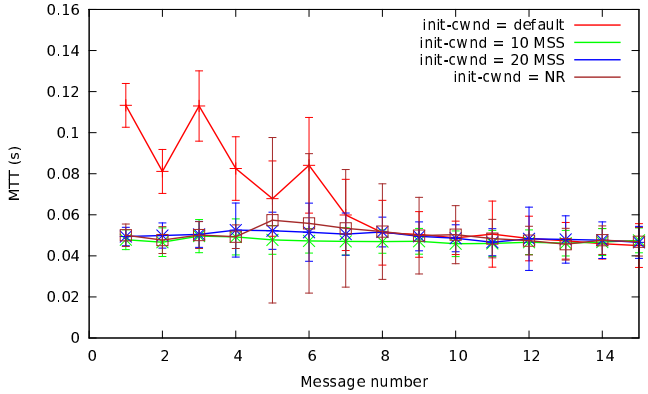


Fig. 7. MTT for one video flow. Handover to a high capacity network with two competing flows.

It is seen that a significant handover delay appeared in those cases the *init_cwnd* was set to the default value of about 3 MSS, but that this delay decreased considerably already as the *init_cwnd* was increased to 10 segments. The reason to this was primarily due to the RTT. An RTT of 40 ms made the network respond to the sender about successful transmission after the generation of about 2-3 messages, or about 8-12 MSS, i.e., the size of in the *init_cwnd*. Secondly, the bandwidth required by the video flows in the scenario was only a small fraction of the fair share of the capacity on the target path.

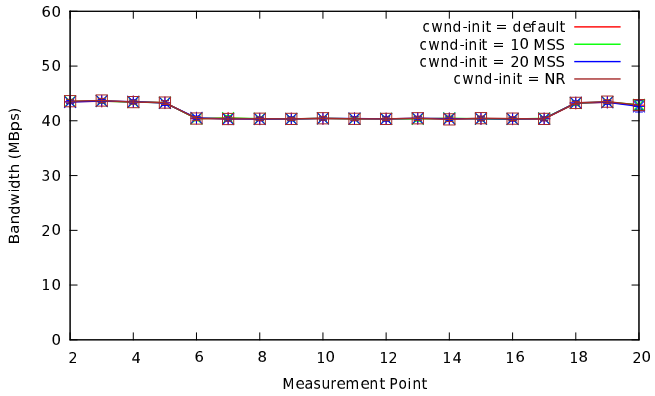


Fig. 8. Throughput for competing traffic. Handover of two video flows to a high capacity network with two competing flows.

Approximately the same results were obtained when handing over two video sessions to a target path where two TCP flows were present. Also for this traffic scenario, we monitored the competing traffic. Figure 8 shows the total capacity used by the two competing TCP flows as the mSCTP based flows were handed over. In the same way as in the *lc* scenario, it is evident that the impact on competing traffic was not affected by an increased *init_cwnd*.

 ISBN: 978-1-61208-229-5

Similar results were obtained in the experiments where five mSCTP flows were handed over to an *hc* target path with two competing TCP flows; a startup delay was observed in those cases the default *init_cwnd* was used, but this delay disappeared when the *init_cwnd* was increased to 10 segments. The result for one of the mSCTP flows in the scenario with 5 mSCTP flows being handed over to a target path with 2 competing flows present are seen in Figure 9[4].

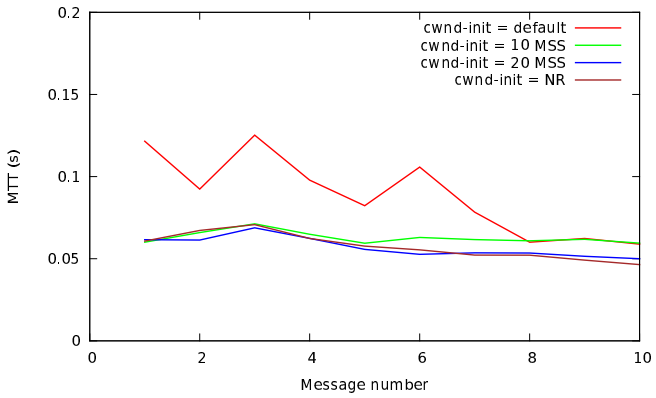


Fig. 9. MTT for one video flow. Handover of five video flows to a high capacity network with two competing flows.

Again, it is seen that by increasing the *init_cwnd* from the default size up to 10 segments, the startup delay after handover was close to zero. When 10 competing flows were sent on the handover target path, the number of lost messages during startup on the target path made the increased *init_cwnd* have no impact on the MTT. When analyzing the quite aggressive competing traffic in the scenario above, we saw no impact on the bandwidth available for the competing traffic from increasing the *init_cwnd*.

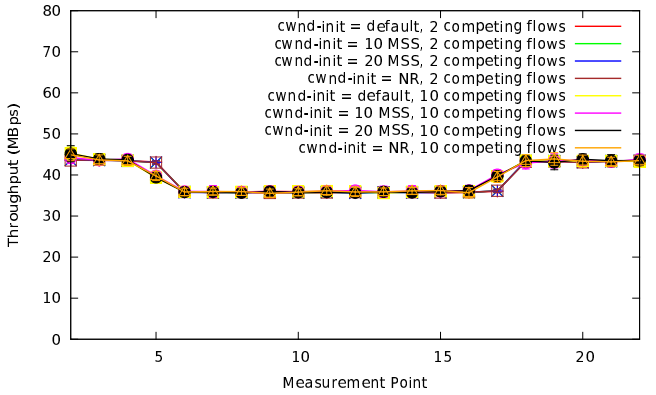


Fig. 10. Throughput for competing traffic. Handover of five video flows to a high capacity network with two vs. 10 competing flows.

Figure 10 shows the capacity used by the competing traffic in cases there were 2 and 10 competing flows present. From the figure we can see no negative impact from increasing the *init_cwnd* in any of the scenarios.

[4]Figure 9 shows the results for mSCTP flow number one out of the five. The results for the other flows look similar to the one shown in the figure.

V. Related Work

The congestion control, introduced in TCP to eliminate the risk of congestion collapses in the network, has over time changed to focus on the challenge of efficient usage of available resources in different types of networks. Several incarnations of the congestion control have as of today been proposed, implemented, deployed in different operating systems. A comprehensive survey of these different congestion control mechanisms, developed for TCP, but also applicable to SCTP, is found in [19].

Dukkipati et al. [20], [21] have recently put forward arguments for an increased *init_cwnd* in TCP. In [21], they study the effects of using an increased *init_cwnd* for Web transactions against geographically spread out data centers. In their study, they suggested that an increased *init_cwnd* could result in significantly decreased latencies for this type of transactions. Our work complements and extends their work by considering the effects on real-time traffic to improve vertical handover for mSCTP.

Previous work on mitigating the effects of slow start during a vertical handover in mSCTP includes SCTP EFC [22] and SHOP [11]. SCTP EFC is a congestion control scheme for mSCTP to be used during handover. The basic idea behind this approach is to store the congestion control parameters of the current primary path when the data flow experiences retransmission timeouts or fast retransmits, i.e., events that typically precede a handover. Later, when a handover takes place, provided the stored parameters has not become obsolete, mSCTP starts out on the handover target path with the stored congestion control parameters. Since mSCTP with EFC starts out on the handover target path with the congestion window size it had on the source path before the handover, it assumes that the network conditions are the same on this path, something that we consider a fairly dangerous assumption.

In SHOP [11], a packet-pair scheme is used to estimate the available bandwidth on the handover target path. On the basis of this estimate, SHOP configures mSCTP's *init_cwnd* and slow-start threshold appropriately. In comparison to SHOP, our proposal with a fixed, increased *init_cwnd* might seem simple and rigid. However, it should be noted that several works have highlighted several problems inherent with measuring available bandwidth [23], [24] – not least by using a packet-pair scheme. Moreover, the idea with an increased *init_cwnd*, is not to increase it up to the available bandwidth, but to set it to a bandwidth that the majority of networks are able to accommodate.

Several other studies have been done on using mSCTP for vertical handover. For example, Koh et al. suggested some tuning guidelines to improve the mSCTP handover performance [25], and demonstrated how it would be possible to integrate mSCTP with MIP [9]. Other works include Cellular SCTP (cSCTP) [26] and SIGMA [27]. cSCTP builds upon mSCTP but differs from it, in that during a handover packets

 ISBN: 978-1-61208-229-5

are duplicated, and transmitted on both the handover source and target paths. Similar to cSCTP, SIGMA uses both the source and target paths during a handover. The SIGMA architecture has in a later work called ECHO [28] been improved to enable QoS-aware handovers.

VI. Conclusion and future work

In this paper, we have studied the impact of an increased initial congestion window for real-time traffic after a vertical handover to a target path where competing TCP traffic is present. The study has been conducted with network parameter settings intended to be representative of wireless 3G and 4G cellular networks.

We have seen that the flow or flows handed over to the target path may significantly benefit from an increased initial congestion window in terms of lower message transfer times. Since the traffic present on the target path does not always back off appropriately, the requirement is that the fair share of the capacity on the target path is far above the capacity required by the sending application.

Additionally, the results show no negative impact on the competing TCP flows by an increased initial congestion window. This fact is important, since a more aggressive startup mechanism should not penalize other traffic.

The work in this area will proceed by integrating the option of a larger initial congestion window in an Android platform of ours, to be able to verify the impact of this altered startup mechanism in real vertical handover scenarios.

Acknowledgment

The work has been supported by grants from VINNOVA, the Swedish Governmental Agency for Innovation Systems.

References

[1] R. Stewart, "Stream Control Transmission Protocol," RFC 4960, Sep. 2007.

[2] R. Stewart, Q. Xie, M. Tuexen, S. Maruyama, and M. Kozuka, "Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration," RFC 5061, Sep. 2007.

[3] J. Eklund, K.-J. Grinnemo, A. Brunstrom, G. Cheimonidis, and Y. Ismailov, "Impact of Slow Start on SCTP Handover Performance," in *Proc. of 20th International Conference on Computer Communications and Networks (ICCCN)*, Maui, HI, USA, Aug. 2011, pp. 1–7.

[4] J. Eklund, K.-J. Grinnemo, and A. Brunstrom, "On the Use of an Increased Initial Congestion Window to Improve mSCTP Handover Performance," in *Proc. of 26th IEEE International Conference on Advanced Information Networking and Applications (AINA)*, Fukuoka, Japan, Mar. 2012, pp. 1101–1106.

[5] T. Dreibholz, E. Rathgeb, I. Ruengeler, R. Seggelmann, M. Tuexen, and R. Stewart, "Stream control transmission protocol: Past, current, and future standardization activities," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 82–88, Apr. 2011.

[6] R. Stewart, M. Ramalho, Q. Xie, M. Tuexen, and P. Conrad, "Stream control transmission protocol (SCTP) partially reliable extension," RFC 3758, May 2004.

[7] M. Tuexen, R. Stewart, P. Lei, and E. Rescorla, "Authenticated chunks for the stream control transmission protocol (SCTP)," RFC 4895, Aug. 2007.

[8] M. Mathis, J. Mahdavi, S. Floyd, and A. Romanov, "TCP selective acknowledgement options," RFC 2018, Oct. 1996.

[9] S. J. Koh, H. Y. Jung, and J. H. Min, "Transport layer internet mobility based on mSCTP," in *IEEE 6th International Conference on Advanced Communication Technology (ICACT)*, Korea, sep 2004, pp. 329–333.

[10] Y. Kim and S. Lee, "mSCTP-based handover scheme for vehicular networks," *IEEE Communications Letters*, vol. 15, no. 8, pp. 828–830, Aug. 2011.

[11] K. Zheng, M. Liu, Z.-C. Li, and G. Xu, "SHOP: An integrated scheme for SCTP handover optimization in multihomed environments," in *IEEE Global Telecommunications Conference (GLOBECOM)*, New Orleans, Lousiana, USA, Dec. 2011, pp. 1–5.

[12] P. Soderman, K.-J. Grinnemo, Cheimonidis, Y. Ismailov, and A. Brunstrom, "An SCTP-based Mobility Management Framework for Smartphones and Tablets," in *Proc. of 26th IEEE International Conference on Advanced Information Networking and Applications (AINA)*, Fukuoka, Japan, Mar. 2012, pp. 1107–1112.

[13] J. Garcia, E. Conchon, T. Perennou, and A. Brunstrom, "KauNet: Improving Reproducibility for Wireless and Mobile Research," in *Proc. of 1st international workshop on system evaluation for mobile platforms(MobiEval07)*, San Juan, Puerto Rico, Jun. 2007, pp. 21–26.

[14] "Dummynet homepage," Jun 2012. [Online]. Available: info.iet.unipi.it/luigi/dummynet/

[15] G. V. Auwera, P. T. David, and M. Reisslein, "Traffic and quality characterization of single-layer video streams encoded with H.264/AVC advanced video coding standard and scalable video coding extension," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 698–718, 2008.

[16] P. Seeling, M. Reisslein, and B. Kulapala, "Network performance evaluation using frame size and quality traces of single-layer and two-layer video: A tutorial," *IEEE Communications Surveys and Tutorials*, vol. 6, no. 2, pp. 58–78, 2004.

[17] "Iperf homepage," Jun 2012. [Online]. Available: http://sourceforge.net/projects/iperf/

[18] A. Vishwanath, V. Sivaraman, and M. Thottan, "Perspectives on router buffer sizing: recent results and open problems," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 2, pp. 34–39, Mar. 2009.

[19] A. Afanasyev, N. Tilley, P. Reiher, and L. Kleinrock, "Host-to-host congestion control for tcp," *Communications Surveys Tutorials, IEEE*, vol. 12, no. 3, pp. 304–342, 2010.

[20] J. Chu, N. Dukkipati, Y. Cheng, and M. Mathis, "Increasing TCP's Initial Window," IETF Internet draft, work in progress, Oct. 2011.

[21] N. Dukkipati, T. Refice, Y. Cheng, J. J. Chu, T. Herbert, A. Agarwal, A. A. Jain, and S. Natalia, "An argument for increasing TCP's initial congestion window," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 3, pp. 26–33, 2010.

[22] K. Lee, S. Nam, and B. Mun, "SCTP efficient flow control during handover," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, Las Vegas, Nevada, USA, Apr. 2006, pp. 69–73.

[23] C. Dovriolis, P. Ramanathan, and D. Moore, "What do packet dispersion techniques measure?" in *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Anchorage, Alaska, USA, Apr. 2001, pp. 905–914.

[24] K. Lai and M. Baker, "Nettimer: A tool for measuring bottleneck link bandwidth," in *Proc. USENIX Symposium on Internet Technologies and Systems*, Boston, Massachusetts, USA, Mar. 2001, pp. 123–134.

[25] S. J. Koh, M. J. Chang, and M. Lee, "mSCTP for soft handover in transport layer," *IEEE Communications Letters*, vol. 8, no. 3, pp. 189–191, Mar. 2004.

[26] I. Aydin, W. Seok, and C.-C. Shen, "Cellular SCTP: a transport-layer approach to Internet mobility," in *The 12th International Conference on Computer Communications and Networks (ICCCN)*, Dallas, Texas, USA, Oct. 2003, pp. 285–290.

[27] S. Fu, L. Ma, M. Atiquzzaman, and Y.-J. Lee, "Architecture and Performance of SIGMA: A Seamless Mobility Architecture for Data Networks," in *IEEE International Conference on Communications (ICC)*, Seoul, Korea, May 2005, pp. 3249–3253.

[28] J. Fitzpatrick, S. Murphy, M. Atiquzzaman, and J. Murphy, "ECHO: A quality of service based endpoint centric handover scheme for VoIP," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*. Las Vegas, Nevada, USA: IEEE, Mar. 2008, pp. 2777–2782.

 ISBN: 978-1-61208-229-5

Reconfiguration of Legacy Software Artifacts on Resource Constraint Smart Cards

Daniel Baldin, Stefan Grösbrink, Simon Oberthür

Design of Distributed Embedded Systems
Heinz Nixdorf Institute, University of Paderborn
Fuerstenallee 11, D-33102 Paderborn, Germany
dbaldin@upb.de, stefan.groesbrink@hni.upb.de, oberthuer@upb.de

***Abstract*—Today's adaptable architectures require the support of configurability and adaptability at design level. However, modern software products are often constructed out of reusable but non-adaptable legacy software artifacts (e.g., libraries) to meet early time-to-market requirements. Thus, modern adaptable architectures are rarely used in commercial applications, because the effort to add adaptability to the reused software artifacts is just too high. In this paper, we describe a methodology to semi-automatically use existing binaries in a reconfigurable manner. It is based on building the annotated control flow graph to identify and extract code on static basic block level depending on different execution requirements given as a set of constraints. This allows for adaptation of binaries after compile time without the use of the corresponding source code. We propose a way of adding additional reconfiguration support to these binary objects. With this approach, reconfiguration can be added with a low effort to non-adaptive software.**

Keywords*—*Reconfiguration; Legacy Software; Smart Cards

I. INTRODUCTION

Software developers often use existing pre-compiled software libraries for various reasons. One reason may be the reduced development time by using third party libraries. Sometimes the use of third party hardware components may also require the use of so called board support packages. In other cases the reason for using pre-compiled libraries may even be as simple as missing source code or documentation. While using these libraries greatly eases the development of new software products, they may also be a source of problems in very resource constraint embedded systems.

Runtime reconfiguration can be the enabling technology for these kind of embedded systems such as Smart Cards by allowing temporarily unused functionalities to be replaced by currently needed functionalities. However the use of pre-compiled third party libraries limits the reconfigurability of the system. State of the art approaches try to solve this problem by wrapping the whole legacy library into a reconfiguration component, leading to a huge waste of memory. Thus, if we want to efficiently use existing libraries inside a reconfigurable system, which cannot be modified at source code level and contain huge amounts of unused or rarely used code, a new approach is required.

In this paper, we introduce a methodology which semi-automatically adds reconfigurability to binary objects using a set of constraints which specify reconfiguration points by high level expressions. The approach is based on creating an annotated control flow graph of the binary on static basic block level and requires only minimal source code information. Specifically, we analyze method signatures to identify higher level expressions that are used for the identification of reconfiguration entry points of the software. The availability of method signatures is only a small restriction since even proprietary libraries include header files containing structure and method signatures describing the Application Program Interface (API) of the library. If this is not be the case, the entire library would not be usable by any higher level programming language as the interfaces would be unknown.

The remaining paper covers the overall methodology of our reconfiguration framework implemented for the ARMv4 Instruction Set Architecture (ISA) in detail, starting with the basic techniques used, followed by the component model, the identification of components, optimizations and concluding with an explanation on the modifications of the original system. Our case study, the evaluation section is based on, focuses on an Internet-Protocol Stack library for an ARM powered Smart Card containing protocol implementations for IPv4 [1], IPv6 [2], TCP [3], UDP [4] and TLS [5]. The scenario contains a web-server which offers communication ports using all of these protocols of the library. However, at runtime not all protocols are used at the same time, which makes it interesting to use the corresponding protocols as reconfiguration components. The paper concludes with related work and outlook.

II. METHODOLOGY

Our approach allows the use of code from legacy libraries as well as fine-granular reconfiguration without the drawbacks of current state of the art approaches. Common approaches either do not allow legacy libraries to be used or simply wrap the complete legacy library into one huge component. This however is not practical for very resource constraint systems. Libraries are typically not given as high level code which might be rewritten for reconfiguration support. Thus, a low level method to extract components out of these libraries and to add reconfiguration support to them is needed. Forcing the

 ISBN: 978-1-61208-229-5

user to do this manually is something that is highly undesirable as well as often impractical as the expert knowledge required to do this cannot be assumed to be available. With this in mind the approach proposed in this paper focuses on the following requirements:

- Usability: Converting parts of the legacy code into reconfigurable components shall be supported by an automated tool that supports to configure the system parameters.
- Run-Time Efficiency: Component loading and replacement shall be as simple as possible without any need of linking the components at run-time. The execution overhead at runtime shall be kept as small as possible.
- Correctness: The semantics of the legacy code must not be changed.

All of these requirements are covered by the approach described in the next sections. The usability is improved by the use of an automatic binary analysis step in combination with the possibility of allowing the user to specify components with high level constraints. Run-time efficiency is achieved by minimizing the overhead of the run-time reconfiguration approach by statically resolving dependencies and by optimizing the components based on parameters as memory and binary overhead, as well as the worst case number of reconfigurations at runtime. The correctness is ensured by the use of instrumentation code which does not change the context of the application. The overall approach is depicted in Figure 1. The approach uses the binary objects, a reconfiguration manager including a replacement policy and a configuration file as the input. The first step is the *binary analysis* of the legacy objects which is covered in the next section. Some of the steps of the approach are ISA specific. In this paper, we will focus on the ARMv4 ISA as our evaluation platform is an ARMv4 powered Smart Card.

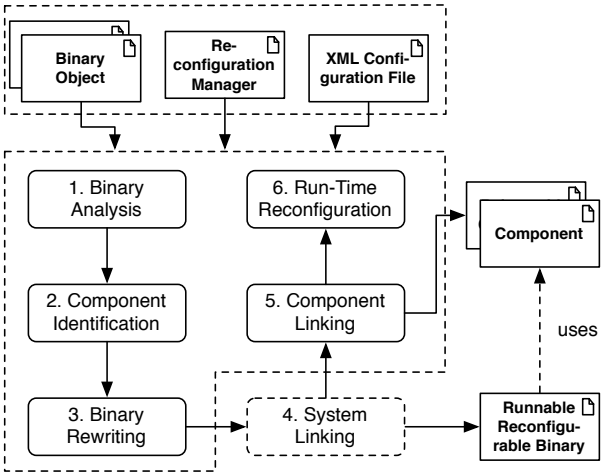


Fig. 1. Steps of the Reconfiguration Approach

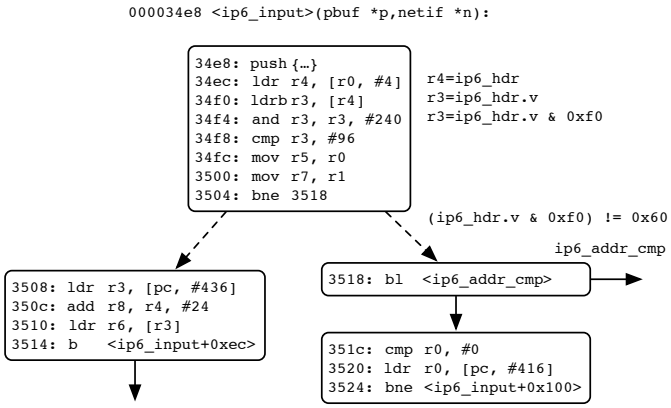


Fig. 2. Parts of the annotated control flow graph of the ip6_input method.

A. Binary Analysis

By disassembling the binary code the static basic blocks and the control flow between these blocks of the program are identified. A static basic block is a sequence of instructions that has exactly one entry point and one exit point. We use the basic block as the smallest representation unit since it describes a linear flow of instructions. A non-linear control flow appears only at the end of a basic block. Each instruction that is a target of a branch instruction defines a new basic block. In general, every program can be uniquely partitioned into a set of non-overlapping static basic blocks.

Figure 2 depicts the first four basic blocks of the disassembled `ip6_input` method. Using these blocks a graph representing the possible control flow of the processor as seen in the Figure is derived. This graph is called the Control Flow Graph (CFG). Each node defines a basic block and the edges represent conditional control flow (dashed edges) and unconditional control flow (solid egdes) between these blocks. Each control flow edge models a dependency between the basic blocks, as reaching one basic block means that we may also reach the successors of it.

The analysis of binary code is a non-trivial task. While disassembling and interpreting binary files, one may encounter several problems as, e.g., the Code Discovery Problem. Many ISAs allow binary data to be mixed up with executable instructions and vice versa. Not being able to distinguish between instructions and data may invalidate the extraction process since some control flows may not be discovered or data may be misinterpreted. However, for our evaluation platform this problem does not exist, since the ARM Embedded Applications Binary Interface (EABI) forces all EABI conform Embedded Linker File (ELF) object files to provide information on all occurrences of data and instruction blocks by special mapping symbols inside the symbol table (see Section 4.6.5 in [6] for the symbol definition).

Another problem with control flow detection arises if indirect control flow instructions are used inside the binary. Most of the indirect control flows are due to jump tables that are generated by the compiler to speed up switch/case statements. The targets of these jumps can be computed with high precision as it was shown by Cifuentes et al. [7]. Other sources of indirect control flows are method pointers, available in most high-level languages, e.g., to implement inheritance

 ISBN: 978-1-61208-229-5

or to realize dynamic program behavior. The targets of these kind of indirect control flows are very hard to compute and to the best of our knowledge no approach exists which can guarantee the precise detection of all targets. However, using the approach proposed by B. Sutter et al. [8] we may overestimate the set of jump targets by introducing a so called *hell node*. The estimation uses the complete set of relocatable symbols, which is the union of all relocatable symbols of all object files, as the target for every indirect jump that can not be resolved. The result may not be as tight as possible but it ensures the correctness of the following reconfiguration process.

In the next step, the edges of the graph are annotated using the common approach of forward substitution. As described by previous work of Cifuentes et al. [9] [10], we derive complex expressions from low level expressions, which in our case are the assembler instructions of the ARM binary. For assembly code one can express the contents of a register r in terms of a set a_k at instruction i as $r = f_1(\{a_k\}, i)$. If the definition at instruction i is the unique definition of a register r that reaches an instruction j along all paths in the program, without any of the registers a_k being redefined, one can forward substitute the register definition at instruction j with $s = f_2(\{r\}, j)$, resulting in:

$$s = f_2(\{f_1(\{a_k\}, i)\}, j)$$

Using these expressions it is also possible to annotate the edges of the control flow graph with constraints that need to be fulfilled for the edge to be taken. However, these expressions consist only of very low level type of operations and resources as, e.g., binary operations and registers. In order to allow these expressions to be used by some developer, it is important to derive as many high level programming language expressions out of the low level expressions as possible. We developed a binary analysis framework [11] which utilizes the high level information stored inside header files to extract type information on the input parameters and global variables of the binary objects. Using global data flow analysis techniques it is possible to annotate parts of the control flow graph with high level constraints based on the input parameters of the binary objects. The result of this analysis has partially been annotated next to the corresponding instruction in Figure 2. However, detecting access to high level data structures is not trivial as an unlimited number of access possibilities to these data structures can be generated by a compiler. In consequence, an expression normalization step as described in [11] is mandatory to allow a meaningful and usable annotation of the binary code.

B. Component Model

After the binary analysis, the binary objects are represented by its CFG $G = (N, E)$, with N being the set of nodes (static basic blocks), E the set of edges and $S \subseteq N$ the set of start nodes (entry points). The function $c : E \rightarrow C$, with C being the set of all constraints, matches every edge to its specific constraint that has been calculated in the previous step. We now need to identify sets of basic blocks inside the CFG which we may use as components inside the reconfiguration process. With a specific input language, the user is able to specify constraints on variables or method parameters used inside the binary objects. An example for such an constraint is given in Listing 1, which has been used for our evaluation scenario. Constraints are either specified for API functions or globally visible symbols. The former ones can contain arbitrary binary operations as the constraints in line two, four and six of Listing 1. The latter ones directly define reconfiguration points as the constraint in line eight. The constraint set is part of the configuration XML file, which is given as an input parameter to the reconfiguration framework, as shown in Figure 1.

```
1 [ip4_input]
2   (ip4_header._ttl_proto & 0x00ff) != 0x06
3 [ip6_input]
4   ip6_hdr.nexthdr != 0x6
5 [ethernet_input]
6   eth_hdr.type != 0x86dd
7 [tls_input]
8   @tls_input
```

Listing 1. Constraints set masking the TCP, IPv4 and IPv6 support as components

For every edge $e \in E$ of the CFG, we then combine the edge constraint $c(e)$ and the corresponding constraint of the user by a logical *and* operation. Our framework uses a constraint solver, which tries to check the satisfiability of the expressions. The set of edges, for which the expression is unsatisfiable, defines the set $R \subseteq E$ that we call set of reconfiguration edges. For unsatisfiable expressions, there exists no assignment of values that satisfy the expression. Our framework implementation currently allows different constraint solvers to be used. For our evaluations, we utilized the STP Constraint Solver [12].

Using the reconfiguration edges of set R, it is now possible to define some important sets of basic blocks, which will be used for our component model throughout the rest of the paper.

Definition II.1 (Mandatory Set)**:**
We define the set M of nodes that can be reached from the start nodes S without taking any reconfiguration edge as $M = \{n \in N : \exists w = (w_1, w_2, ..., w_n), w_1 \in S \wedge (w_i, w_{i+1}) \in E \setminus R \wedge w_n = n\}$. This set defines the set of basic blocks that we call the Mandatory Set.

Definition II.2 (Intermediate Components)**:**
For every reconfiguration edge $r_i \in R$ with $r_i = (n_{i1}, n_{i2})$ we define the set N_{r_i} of nodes that can be reached over the reconfiguration edge r_i without reaching a node that is mandatory (inside set M): $N_{r_i} = \{k \in N : \exists w = (w_1, w_2, ..., w_n) \wedge w_1 = n_{i2} \wedge (w_j, w_{j+1}) \in E \wedge w_j \notin M \wedge w_n = k\}$. We call these sets Intermediate Components.

Both sets can easily be computed by using a depth first search starting at the start nodes S for finding the Mandatory Set, or at the nodes $\{n_{i2}\}$ with $r_i \in R, r_i = (n_{i1}, n_{i2})$ for finding the Intermediate Components respectively using

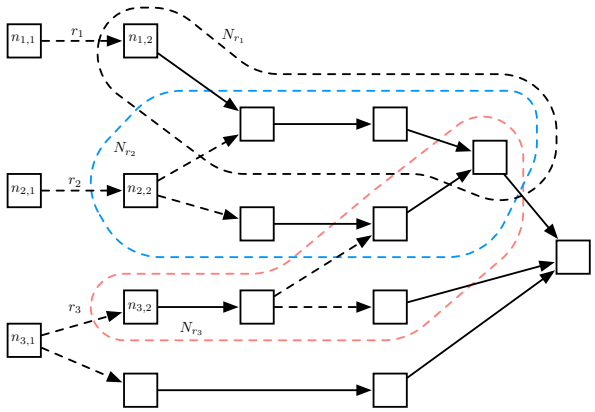


Fig. 3. Example Intermediate Component sets based on definition II.2

the restrictions inside the corresponding definition. As the name Intermediate Component suggests, these sets are used intermediately and form the basis for the final components used inside the reconfiguration process.

C. Component Identification

The initially computed sets $N_{r_1}, ..., N_{r_n}$ already describe which kind of functionality can be executed when the corresponding edge $r_i \in R$ is taken. These sets however may be ambiguous as seen in Figure 3. Using these sets of intermediate components without further refinement could create duplicate code segments, which is highly undesired. In order to resolve all ambiguities, Algorithm 1 can be used to generate distinct components that, while depending on each other, can be used efficiently inside the reconfiguration process. The basic idea is to generate all possible intersections as longs as there exist ambiguous sets of basic blocks. In each intersection iteration (see line five of Algorithm 1), all possible intersections of the current working set S are calculated. Redundant intersections or empty intersections are not stored. At the end of the intersection step all basic blocks contained inside any of the intersection sets K_i are removed from the configurations inside the working set S. The created sets K_i then define the working set for the next intersection step. The iteration ends if the working set contains only one or no set anymore as there exists no possible new intersection that may be computed. The iterations (given by the while loop in line five) of the algorithm on the example graph of Figure 3 can be seen in Figure 4 .

Using Algorithm 1, it is possible to split up the Intermediate Components into single distinct sets of basic blocks. However, this will introduce dependencies between each of the sets, which are defined by the control flow edges between them.

Definition II.3 (Dependency)**:**
Given two components S_i, S_j, if there exists an edge $e = (n_1, n_2)$ with $n_1 \in S_i, n_2 \notin S_i$ and $n_1 \notin S_j, n_2 \in S_j$ we say S_i *directly depends* on S_j, denoted by $S_i \rightarrow S_j$. If there exists a path $w = (w_1, ..., w_n)$ with $w_1 \in S_i, w_2, ..., w_{n-1} \in M, w_n \in S_j$ we say S_i *depends* on S_j, denoted by $S_i \rightsquigarrow S_j$.

The corresponding direct dependencies between components inside the example graph can be seen in Figure 4. The "direct" dependency graph of the initial components can never contain loops due to the construction of it. However, the dependency graph may contain loops as control flow from components may happen to the mandatory set and back. Using the dependency graph, it is possible to estimate the runtime overhead for different paths of the application if the reconfiguration time is known. The execution time of the code inside the components does not change as the application code is not changed. The only source of execution time changes inside the components may result from different caching and pipeline effects which we do not consider.

Algorithm 1 Component Identification

1: **procedure** GENERATECOMPS($N_{r_1}, ..., N_{r_n}$) ▷ Input: N_{r_i} of definition II.2
2: Set $S \leftarrow \{N_{r_1}, ..., N_{r_n}\}$
3: Set $K \leftarrow \{\}$
4: Set $R \leftarrow \{\}$ ▷ The set of output components
5: **while** $|S| > 1$ **do**
6: **for all** $S_i, S_j \in S, S_i \neq S_j$ **do**
7: $T \leftarrow S_i \cap S_j$
8: **if** $T \notin K \wedge T \neq \{\}$ **then**
9: $K \leftarrow K \cup \{T\}$ ▷ Add the set T to K
10: **end if**
11: **end for**
12: **for all** $S_i \in S$ **do**
13: $S_i \leftarrow S_i \setminus \left(\bigcup_{K_i \in K} K_i \right)$ ▷ Remove all sets in K from S_i
14: $R \leftarrow R \cup \{S_i\}$ ▷ Add a new layer of components
15: **end for**
16: $S \leftarrow K$
17: $K \leftarrow \{\}$
18: **end while**
19: **return** R
end procedure

D. Optimal Component Size

The extracted components may now be used for reconfiguration. However, using the components in this state may be far from optimal if we consider factors as runtime overhead, binary overhead or memory fragmentation. Especially in Smart Cards, innately being very resource constraint systems, these overheads need to be kept as small as possible. Many Smart Cards use flash memory for the non-volatile memory space. Most of them also execute applications directly out of flash memory. The use of flash memory inherently raises the demand of reducing the amount of flash page writes at runtime as the lifetime of the memory page is limited by a certain number of erase/write operations. One objective optimization function would thus try to minimize the number of such operations. We are currently only focusing on components which are

 ISBN: 978-1-61208-229-5

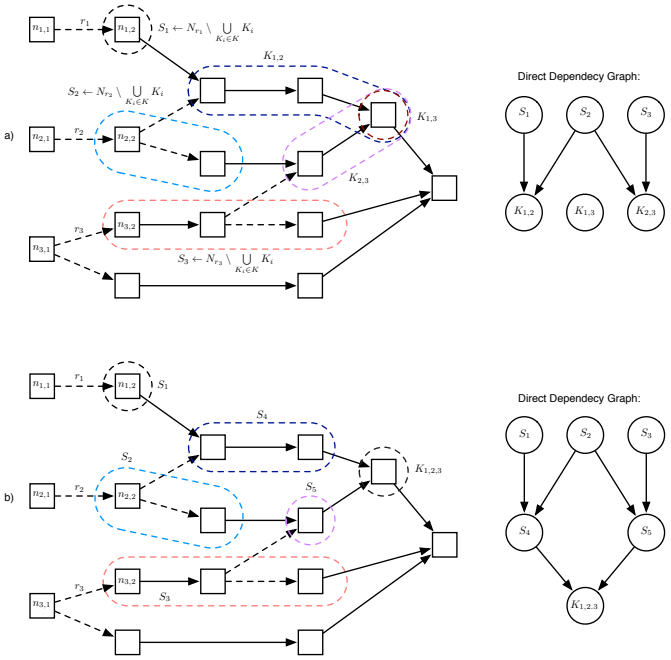


Fig. 4. The iterations of Algorithm 1 on the example graph.

bigger than the minimal flash page size. In order to reduce the number of write operations during reconfiguration, only one component may be placed in one flash page. In addition, the component size is kept as close as possible to multiples of the page size in order to reduce the memory fragmentation at runtime.

This problem is solved by splitting up the components $S_1, .., S_n$ into smaller components. The components $S_1, .., S_n$ with sizes $s_1, .., s_n$ exceeding a size s may be split up into multiple components of size s and one component of size s_i modulo s. This can be done in several ways. One may simply use the linear binary object layout and split the basic blocks at the corresponding positions or one may use a reordering approach which tries to place strongly connected basic blocks close together, just like a compiler would do.

On the one hand, splitting the components will introduce new dependencies resulting in a higher number of reconfiguration edges. On the other hand, this changes the memory fragmentation introduced by each component. Depending on the target system, these attributes will be more or less important. However, an optimal solution to this problem often does not exist because properties as runtime overhead and memory fragmentation are, typically, in contrast to each other and highly dependent on the application. In the next step we thus perform a design space exploration.

Definition II.4 (Design Points)**:**
We define the set of design Points $D_{(m,s)} = (D_B, D_M, D_R)$ as follows: for possible component sizes $s = x \cdot F_{min}$ (with F_{min} being the smallest page size) and total reconfiguration memory space $m = r \cdot s$, the values for the binary overhead D_B and memory fragmentation D_M as well as the worst case number of reconfigurations D_R are calculated.

Our approach of solving this problem is to do a multi-objective optimization by calculating the Pareto optimal points using the Greaf-Younes algorithm [13] with backward iteration over the set of all design points $D_{(m,s)} = (D_B, D_M, D_R)$. Given the total amount of reconfiguration space m it is possible to store r components of size s before runtime replacement (based on some function $f_{replace}$) starts. Using this definition the design space values are calculated in the following way:

- *Binary Overhead D_B:*
  The Binary Overhead is defined as the median percentage based increase of the component size due to added instrumentation code. If the number of jumps / references between components increases, the amount of instrumentation code may increase as well.
- *Memory Fragmentation D_M:*
  For every component size s we sum up the r highest fragmentations of components as this yields the worst case situation with the highest amount of wasted memory.
- *Worst Case Number of Reconfigurations D_R:*
  The worst case number of reconfigurations is the maximum number of reconfigurations needed to execute any possible path inside the context sensitive control flow graph. The path analysis is context-sensitive and done using a Depth First Search approach. Iterating over all possible context-sensitive paths is in general infeasible for even small graphs, however, we use a branch and bound based algorithm to avoid traversing any path which can not increase the worst case reconfiguration number. The value can be calculated more easily by restricting candidate paths to start at the incoming edges of components and to follow a path inside the dependency graph. The function $f_{replace}$ is used to simulate the runtime replacement. If a loop over more than r components is encountered the design point is removed from the design space as we are not able to estimate the number of reconfigurations needed to execute the loop at runtime without knowing loop boundaries.

The final component size s and reconfiguration space m may then be chosen from the set of Pareto optimal points either by the system designer or by using a user defined rating function. The evaluation section will cover an example for determining the optimal component size for our evaluation scenario. It is important to note that increasing the reconfiguration space m will not always lead to better design space points as discussed in the Section Evaluation (IV).

E. Binary Rewriting

As the components are given by the previous steps, we now have to add reconfiguration support to them. The extracted sets of basic blocks usually contain references to relocatable symbols, which would have been resolved at link time of the binary. Relocatable symbols may reference different kinds of application sections as, e.g., the executable code area of other

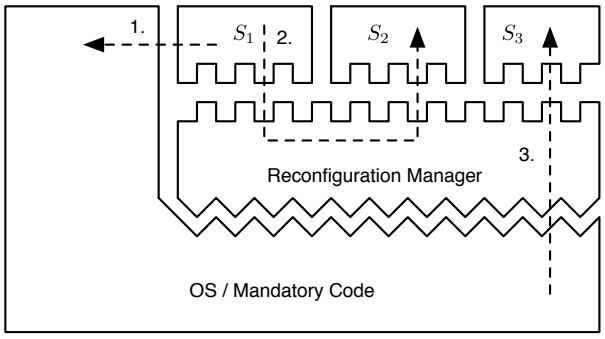


Fig. 5. The reconfiguration architecture and some possible control flows: (1.) control flow from a component to the mandatory set, (2.) control flow from between components, (3). control flow from the mandatory set to a component

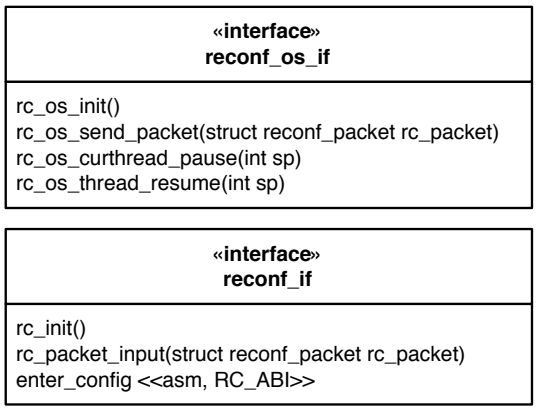


Fig. 6. The interface required/provided by the reconfiguration manager.

components, read-only data or the heap of the component. Relocatable symbols, which reference addresses inside the mandatory set M, are resolved after link time in step *component linking* (see Figure 1). References to other reconfiguration components are replaced by instrumentation code, which adds a call to the reconfiguration manager. Currently, unsupported are references to the data areas of components. The current solution places data areas of the reconfiguration component into the mandatory set, thus allowing the reference to be solved after the mandatory set has been linked.

The instrumentation code is added to the components and placed as close as possible to the corresponding reference in oder to avoid additional overhead of implementing long jumps. The framework also uses the live register information gained by the binary analysis step to use free registers to implement the call to the reconfiguration handler. If all registers are used the context is temporarily stored on the stack. The binary overhead introduced by the instrumentation code may thus vary between four and twenty bytes depending on the free register set and the instruction set (ARM features a 16 bit THUMB and a 32 bit ARM instruction mode). The overhead introduced for a realistic example is discussed in the evaluation section.

The components themselves also need to be rewritten. This is required as references to other basic blocks inside the components may be invalid due to the fact that basic blocks may have been added, removed or changed. Thus, all instructions containing references to other basic blocks inside the component are updated. The same holds true for symbol and relocation entries inside the binary which are used by the linker. The process of modifying these offsets is described inside [11].

III. Run-Time Reconfiguration

At runtime, transparently to the user, the components are exchanged on demand. The architecture depicted in Figure 5 describes the possible control flows and the interfaces involved can be seen in Figure 6. The reconfiguration manager is the central part of the reconfiguration process. In order to work properly, the interface `reconf_os_if` needs to be implemented by the operating system. Among others, it defines a method which is used to send reconfiguration packets to the reconfiguration server, which runs for example on a terminal connected to the smart card. In our example scenario we use a UDP based connection.

Basically, three types of control flows needs to be covered by the system. Control flow going from a component to the mandatory code / OS (see 1. in Figure 5) is already covered by the binary rewriting process. As the mandatory code is not moved inside the physical address space, the corresponding branches are handled by the instrumentation code. The runtime overhead of these control flows is static. Control flow occurring between components (see 2. in Figure 5) involves the reconfiguration manager. Let us consider the components S_1 and S_2 inside the figure. The component S_1 needs to execute some code in component S_2. As the physical memory location of the component changes and/or the component may not be loaded, the reconfiguration manager is called first. The reconfiguration manager provides an assembler routine named `enter_config`, which is automatically called by the instrumentation code added to the components. The assembler routine takes the ID and the offset of the code to be executed inside the component as parameters. If the component is loaded the call is forwarded and the code is executed. This takes around ten assembler instructions on the ARMv4 THUMB ISA. If the component is currently not loaded a reconfiguration request is issued which will use the interface to the operating system to reload the component. The issuing thread context is saved on the stack and stored by the reconfiguration manager to resume the thread upon completion of the reconfiguration. Control flow from the mandatory code to a component (see 3. in Figure 5) involves the same steps as the previous one. It is handled by the same assembler routine and the call to the reconfiguration manager is also automatically added to the mandatory code by the binary rewriting step.

The replacement strategy $f_{replace}$ is implemented inside the reconfiguration manager and used to determine which component will be replaced at runtime. The current implementation uses a *least frequently used* (LFU) replacement function. For

 ISBN: 978-1-61208-229-5

Design Flow Step	Execution Time
Header Analysis	7929 ms
CFG Generation	4988 ms
DF Analysis	33921 ms
Constraint Checking	264 ms
Component Identification	297 ms
Binary Rewriting	963 ms
Component Optimization	6560 ms

TABLE I
EXECUTION TIME OF THE DESIGN FLOW STEPS FOR THE EXAMPLE SCENARIO.

every reconfiguration edge taken a counter is increased for the corresponding component. If replacement takes place the component with the smallest counter is removed. The LFU algorithm was chosen as the implementation overhead of this algorithm is very small.

IV. EVALUATION

This section gives an evaluation of the binary reconfiguration approach.

A. Case Study

Our case study is an Internet-Protocol Stack library for an ARM powered SmartCard containing protocol implementations for the Internet Protocol Version 4 (IPv4), Version 6 (IPv6), the Transmission Control Protocol (TCP), User Datagram Protocol (UDP) and a Transport Layer Security (TLS) implementation. The Smart Card contains a USB interface, which we use to emulate an ethernet connection using the Ethernet Emulation Model (EEM). With this USB connection the Smart Card is connected to a Linux computer, which uses the Ethernet interface to transparently communicate with the device. The scenario contains a smart card web-server, which offers communication ports using all of these protocols of the library. However, at runtime not all protocols are used at the same time which makes it interesting to use the corresponding protocols as reconfiguration components. The complete binary size of all binary objects inside the case study consists of 44246 Bytes.

B. Design Time Overhead

Our binary reconfiguration framework has been implemented in Java and supports the ARMv4 ISA. However, the software architecture allows for the easy addition of support for other ISAs. The reconfiguration process was executed on a Linux computer with a single core 2,8 Ghz Pentium processor. The execution time of the design flow steps can be seen in Table I. The most time consuming part is the Data-Flow (DF) Analysis which annotates the CFG with high level constraints and resolves indirect branches. The Component Optimization step which includes the calculation of the worst case reconfiguration amount for all design points only took about seven seconds to complete. All together the complete execution time of the framework stayed under one minute which is a reasonable time frame.

Component	Component Size	Complete Size	Percentage
S_1 (TLS)	6948	10252	67,7 %
S_2 (IPv6)	1046	2024	51,6 %
S_3 (TCP)	4136	10468	39,5 %

TABLE II
EXTRACTED COMPONENT SIZES IN BYTES

C. Reconfiguration Manager Overhead

As the reconfiguration itself adds new executable code to the original binary, it is very important to keep this additional code as small as possible. The implementation of the reconfiguration manager including the interface implementation and the replacement function added an additional 680 Bytes of code to the application. Inside the example scenario the communication stack of the operating system could be reused resulting in a small reconfiguration manager.

D. Component Extraction

The XML configuration file contained the constraints shown in listing 1, which were passed to the constraint solver with the goal to extract the TCP, IPv6 and TLS components from the application in order to reuse these components inside the reconfiguration process. Line two and four describe a constraint to identify the control flow to the TCP component, line six specifies the control flow to the IPv6 component and the symbol constraint in line eight describes an entry point to the TLS component. Table II shows the size of the extracted components S_i after using Algorithm 1.

Using this simple constraint set, it was possible to extract 68% of the TLS implementation code to be used inside a reconfiguration component. The remaining bytes of the implementation may be extracted with a more sophisticated constraint set as not all control flows are covered by the set of Listing 1. A similar statement holds true for the TCP and IPv6 components in Table II for which the percentage is lower. This is due to the fact that the constraints only restricts control flow from the lower Ethernet packet layer. Control flow from higher layers, as, e.g., the application layer, was not considered by the constraint set. This is, however, possible without restriction.

E. Component Optimization

In the next step, the design points $D_{(m,s)} = (D_B, D_M, D_R)$ have been calculated. The basic blocks have not been reordered and were placed inside a component based on their linear order inside the object files they were taken from. For every combination of component size and reconfiguration space (m, s) the resulting worst case number of reconfigurations, binary overhead and memory fragmentation has been calculated. Table III gives an overview of some of the calculated points sorted by number of reconfigurations as this has been most interesting for our scenario. The Pareto optimal points have been highlighted. All of the highlighted design points are equally good with respect to the pareto optimality. However, depending on the target architecture and the execution scenario one may favor design points with a small numbers of reconfigurations or low fragmentation.

 ISBN: 978-1-61208-229-5

m	s	D_R	D_B	D_M
4096	256	17	28,06252967	530
...	...	...	...	...
4096	1024	15	19,07260033	1654
3072	512	12	21,420201	954
...	...	...	...	...
4096	512	10	21,420201	954
8192	1024	8	19,07260033	1654
4096	2048	7	14,56403317	2134
...	...	...	...	...
2048	2048	7	14,56403317	1372
5120	1280	7	16,03623833	2166
5120	5120	7	12,3342234	3834
6144	1536	6	16,813032	494
6144	2048	6	14,56403317	2694
7168	1792	5	15,73381667	2110
4096	4096	5	13,16168417	3336
6144	3072	5	15,5311505	3490
...	...	...	...	...
6912	6912	4	12,29584317	5626
7168	3584	3	14,45041367	3394
7168	7168	3	12,21908233	5882

TABLE III
CALCULATED DESIGN POINTS $D_{(m,s)} = (D_B, D_M, D_R)$. THE PARETO OPTIMAL ONES ARE HIGHLIGHTED.

The binary overhead for our evaluation example stayed between 30% and 12% (compare the values D_B of Table III), which is a reasonable increase in code size of the components. We used the design point $D_{(2048,2048)}$ for the evaluation, which resulted in a final component size $m = 2048$ bytes and reconfiguration memory size $s = 2048$ bytes. Using this design point limited the number of concurrent components on the Smart Card to one. Thus, each time a dependency edge is taken between two components a reconfiguration needs to take place. The worst case path consisted of seven reconfigurations. Using the UDP connection to the Linux computer running the reconfiguration server the maximum reconfiguration time for one component (2048 Bytes) took 153 ms. A connection request to the web-server was delayed by a median value of 920 ms. This demonstrates that it is possible to run the web-server application on the Smart Card with a memory requirements of 35 KB. This results in a memory saving of approximately 22%. However the memory saving is paired with a much higher run-time of the application.

Interesting to see is that simply increasing the amount of reconfiguration memory space does not always yield better design points. This can be seen by comparing $D_{(4864,4864)}$ with $D_{(2816,5632)}$. Although the latter design point offers a higher amount of reconfiguration space the system parameters are worse. This shows that the design optimization step is mandatory if the system parameters need to be optimized and/or known beforehand.

V. RELATED WORK

Many approaches have been created to solve parts of the goals described in this thesis. Link-Time optimization approaches [14][15] allow binary code to be optimized for speed and memory requirements. This is a valuable technique which already grants huge benefits for software programs. However, it does not solve the general problem if the execution space is still too small for an application to run on an embedded device. It also is not intended as an approach which allows software programs to be adapted at runtime.

Binary Analysis approaches have been used for many reasons for years. Most of the efforts are concerned with analyzing source code which is not available in the contents of this approach. Approaches which analyze binary code are focused on different problems. On the one hand they are used for the link-time optimization described above. On the other hand it is used to cope with security issues of applications [16], [17] quality assurance or compliance testing [18]. Recently Binary Analysis and Binary Rewriting gained popularity inside the research community again. Modern run-time compiler use data flow analysis techniques do to optimizations by using, e.g., trace-scheduling techniques [19][20]. In this approach, we use binary analysis techniques for a different purpose: it is used to gather information on the binary objects, which will enable run-time reconfiguration of binary objects. To the best of our knowledge, there exists no approach which tries to use binary analysis to support software reconfiguration of legacy software systems.

Run-Time Reconfiguration approaches have been proposed for small embedded devices for different goals. It has been shown to be indispensable for some kinds of applications as it allows for re-tasking, fixing bugs, adding functionality or replacing functionality due to memory restrictions. Reconfiguration is supported by some operating systems, particularly often used inside sensor-networks. For example, Agilla [21] or TinyOS [22] support some form of reconfiguration. However the reconfiguration either consist of full binary upgrades (TinyOS) or requires the source code (Agilla), which makes the use of legacy code impossible. Other forms of binary adaptation may be categorized by the following approaches. Whabe et al. [23] propose the creation of adaptable binaries by adding information to the binaries, which may then be used to modify the binary later on. The approach in [24] is based on using new architectures and creating adaptable and reloadable components on source code level. A promising approach has been shown in [25] by creating so called ”delta files”, which contain the byte streams of the adaptations to be made on binary level. However, the delta files are created by compiling the adaptations from source code for the different kinds of configurations. All these approaches have in common that they cannot be used with proprietary libraries that already have been compiled and may not be rebuilt with these kind of information or adaptation support.

VI. CONCLUSION AND FUTURE WORK

In this paper, we demonstrated an approach which allows binary objects as they are contained inside legacy libraries to be used as components inside a reconfigurable system. By using control flow and data flow analysis techniques the approach derives a high level representation of the binaries. Combined with a constraint satisfaction problem solver the system allows the user to easily define components based on

 ISBN: 978-1-61208-229-5

high level constraints. The components are then optimized regarding parameters as number of reconfigurations at runtime, binary overhead due to the added instrumentation code and memory fragmentation. In the end, the approach allows these components to be seamlessly added and removed at runtime. The evaluation showed that it is possible to extract components out of the binary objects. The optimization step derives a Pareto optimal set of design parameters. In the end the legacy objects can be used as components inside a reconfigurable system which can have a lower memory consumption while maintaining the functionality offered by the legacy objects.

Future adaptations may further increase the benefit gained as in the presented work we only used a simple linear basic block scheduling technique inside the components. Using more sophisticated placing algorithms may further decrease the binary and reconfiguration overhead. We may also consider profile based information to better identify components and to identify heavily used program paths. This may allow the design space exploration to find better design points, which will decrease the median number of reconfigurations.

REFERENCES

[1] "RFC 791 - Internet Protocol Version 4." [Online]. Available: http://www.ietf.org/rfc/rfc791.txt

[2] "RFC 2460 - Internet Protocol Version 6." [Online]. Available: http://www.ietf.org/rfc/rfc2460.txt

[3] "RFC 793 - Transmission Control Protocol." [Online]. Available: http://www.ietf.org/rfc/rfc793.txt

[4] "RFC 768 - User Datagram Protocol." [Online]. Available: http://www.ietf.org/rfc/rfc768.txt

[5] "RFC 4346 - Transport Layer Security Version 1.1." [Online]. Available: http://www.ietf.org/rfc/rfc4346.txt

[6] ARM Ltd., "ELF for the ARM Architecture," 2009.

[7] C. Cifuentes and M. V. Emmerik, "Recovery of jump table case statements from binary code," in *Science of Computer Programming*, 1999, pp. 2–3.

[8] B. D. Sutter, B. D. Bus, K. D. Bosschere, P. Keyngnaert, and B. Demoen, "On the static analysis of indirect control transfers in binaries," in *In PDPTA*, 2000, pp. 1013–1019.

[9] C. Cifuentes, "Interprocedural data flow decompilation," *Journal of Programming Languages*, vol. 4, pp. 77–99, 1996.

[10] C. Cifuentes, D. Simon, and A. Fraboulet, "Assembly to high-level language translation," in *In Int. Conf. on Softw. Maint*. IEEE-CS Press, 1998, pp. 228–237.

[11] D. Baldin, S. Groesbrink, and S. Oberthür, "Enabling constraint-based binary reconfiguration by binary analysis," *GSTF Journal on Computing (JoC)*, vol. 1, no. 4, pp. 1–9, January 2012.

[12] V. Ganesh and D. L. Dill, "A decision procedure for bit-vectors and arrays," in *Computer Aided Verification (CAV '07)*. Berlin, Germany: Springer-Verlag, July 2007.

[13] J. Jahn, "Vector optimization, theory, applications and extensions." Springer-Verlag, 2011, p. 345.

[14] D. W. Goodwin, "Interprocedural dataflow analysis in an executable optimizer," 1997.

[15] W. E. Weihl, "Interprocedural data flow analysis in the presence of pointers, procedure variables, and label variables," in *Proceedings of the 7th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, ser. POPL '80. New York, NY, USA: ACM, 1980, pp. 83–94. [Online]. Available: http://doi.acm.org/10.1145/567446.567455

[16] N. Xia, B. Mao, Q. Zeng, and L. Xie, "Efficient and practical control flow monitoring for program security," in *Proceedings of the 11th Asian computing science conference on Advances in computer science: secure software and related issues*, ser. ASIAN'06. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 90–104. [Online]. Available: http://dl.acm.org/citation.cfm?id=1782734.1782742

[17] D. Wagner and D. Dean, "Intrusion detection via static analysis," in *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, ser. SP '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 156–. [Online]. Available: http://dl.acm.org/citation.cfm?id=882495.884434

[18] R. Venkitaraman and G. Gupta, "Static program analysis of embedded executable assembly code," in *Proceedings of the 2004 international conference on Compilers, architecture, and synthesis for embedded systems*, ser. CASES '04. New York, NY, USA: ACM, 2004, pp. 157–166. [Online]. Available: http://doi.acm.org/10.1145/1023833.1023857

[19] N. V. Mujadiya, "Instruction scheduling for vliw processors under variation scenario," in *Proceedings of the 9th international conference on Systems, architectures, modeling and simulation*, ser. SAMOS'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 33–40. [Online]. Available: http://dl.acm.org/citation.cfm?id=1812707.1812717

[20] E. Yardimci and M. Franz, "Mostly static program partitioning of binary executables," *ACM Trans. Program. Lang. Syst.*, vol. 31, no. 5, pp. 17:1–17:46, Jul. 2009. [Online]. Available: http://doi.acm.org/10.1145/1538917.1538918

[21] C.-L. Fok, G.-C. Roman, and C. Lu, "Agilla: A mobile agent middleware for self-adaptive wireless sensor networks," *ACM Trans. Auton. Adapt. Syst.*, vol. 4, no. 3, pp. 16:1–16:26, Jul. 2009. [Online]. Available: http://doi.acm.org/10.1145/1552297.1552299

[22] J. Hill, R. Szewczyk, A. Woo, S. Hollar, D. Culler, and K. Pister, "System architecture directions for networked sensors," *SIGPLAN Not.*, vol. 35, no. 11, pp. 93–104, Nov. 2000. [Online]. Available: http://doi.acm.org/10.1145/356989.356998

[23] R. Wahbe, S. Lucco, and S. L. Graham, "Adaptable binary programs," IN, Tech. Rep., 1994.

[24] S. Kogekar, S. Neema, and X. Koutsoukos, "Dynamic software reconfiguration in sensor networks," in *Proceedings of the 2005 Systems Communications*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 413–420.

[25] R. Keller and U. Hölzle, "Binary component adaptation," in *Proceedings of the 12th European Conference on Object-Oriented Programming*. London, UK: Springer-Verlag, 1998, pp. 307–329.

 ISBN: 978-1-61208-229-5

On-road Wireless Sensor Network for Traffic Surveillance

JaeJun Yoo, DoHyun Kim, KyoungHo Kim
Vehicle/Ship/Defense IT Convergence Division
Electronics and Telecommunications Research Institute
Daejeon, South Korea
{jjryu, dohyun, kkh}@etri.re.kr

JongHyun Park
Robots/Perception Convergence Division
Electronics and Telecommunications Research Institute
Daejeon, South Korea
jhp@etri.re.kr

***Abstract*— In this paper, a traffic surveillance system using wireless sensor networks is introduced. Such traffic surveillance system should satisfy more requirements and overcome harder constraints because wireless network systems have some characteristics such as small hardware resources and less stability of communication. Such requirements include energy-efficient operation, reliable data transmission, and accurate operations. A traffic surveillance system using magnetic sensor network is designed and implemented to satisfy mentioned requirements. The designed traffic surveillance system is tested on indoor and outdoor test-beds.**

Keywords – WSN; Magnetic Sensor; Traffic Surveillance.

I. INTRODUCTION AND PREVIOUS WORKS

Recently, as various sensor network technologies for ubiquitous computing have been being researched actively, the convergence of existing applications and the sensor network technologies has been being a key issue in several industrial and scientific fields. For example, in the Intelligent Transport System field, the convergence of diverse services, research issues and sensor network technologies has been being suggested [1][2].

Some previous works on wireless sensor network based on magnetic sensor nodes [3][4][5] were introduced to adapt the sensor network technologies on real road environments. The previous works focus on each research topics related to wireless sensor network such as energy efficiency, detection accuracy, traffic management, and so on. However, they do not suggest overall and essential requirements to successfully adapt wireless magnetic sensor network on road networks and do not describe their overall system which are the most important for manufacturing sensor nodes.

In this paper, as a kind of the convergence of wireless sensor networks and Telematics/ITS environments, we design and implement a prototype of a whole traffic surveillance system using wireless magnetic sensor networks on roads. For that, we suggest key and essential requirements and overall architecture including sub-components and signal processing mechanisms reflecting the requirements are explained.

This paper is structured as followings. In section 2, we explain the overall design and structure of the traffic surveillance system including system requirements. In section 3, we conduct some experiments and show the results. In section 4, we conclude this paper.

II. ON-ROAD WIRELESS SENSOR NETWORK FOR TRAFFIC SURVEILLANCE

In this section, the overall design and structure of the traffic surveillance system including system requirements will be explained.

A. Key Requirements

A traffic surveillance system using wireless sensor networks should consider the following requirements [6].

- Energy-efficiency on sensor networks
- Accurate vehicle detection
- Reliable and real-time data transmission on wireless communication
- Real-time sensor data processing and vehicle detection
- Efficient and intuitive management of sensor networks and services

There can be some variations on such requirements according to installation environments of the vehicle detection system. For example, some real-time properties, such as the limit time for system responses, can differ from types of real-roads, such as high-way and in-city roads.

B. System Architecture

Figure 1 below shows the overall structure of the designed traffic surveillance system. The traffic surveillance system mainly consists of three parts according to their functions - 1) vehicle detection layer, 2) data management layer, and 3) monitoring & application layer. The vehicle detection layer consist of magnetic sensor nodes installed on surface of real-roads, gate-nodes to relay vehicle detection data, and base-stations to process detection data and to calculate speed of vehicles. More specifically, the magnetic sensor nodes detect vehicles based on variations of magnetic fields of earth caused by moving vehicles. The magnetic sensor nodes also transfer the vehicle detection information to gate-nodes or base-stations with detection time (local ticks). Base-stations receive the detection information from magnetic sensor nodes or gate-nodes, and process them to extract more useful data such as speed of vehicles. The data management layer provides broader services based on information transmitted from base-stations, such as traffic status of some districts. The monitoring and application layer

 ISBN: 978-1-61208-229-5

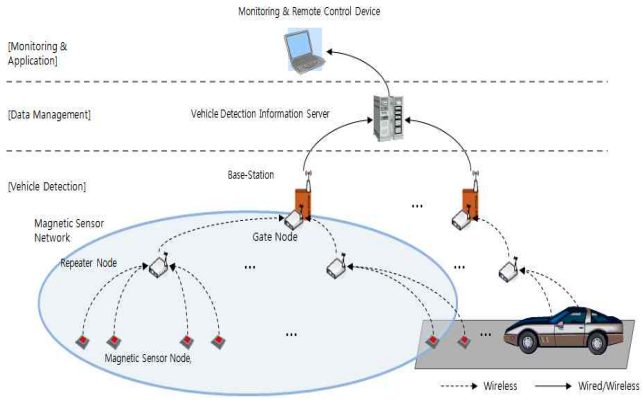


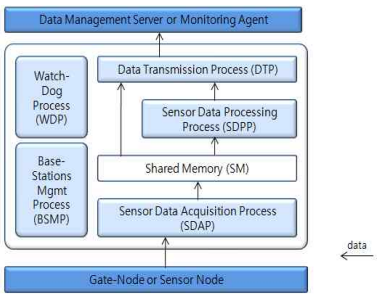
Figure 1. Overall structure of the traffic surveillance system based on magnetic sensor networks

provides more customized services to end-users, such as system monitoring and personalized services.

In this paper, we have focused on the vehicle detection layer, especially wireless sensor network and base-stations, because remains are related to flexible functions according to kinds of traffic-related services.

C. *Base-Stations*

A base-station collects vehicle detection data from magnetic sensor network, processes them to produce more valuable information, for example, speed of a vehicle, and provides them to upper layers. Figure 2 and Table I show the S/W structure and H/W aspects of a base-station.



(a) S/W structure of a base-station

Item	Description
CPU	MPC5200 400MHz (Power PC Core)
Memory (SDRAM)	SDR-SDRAM 128 MB
Memory (Flash)	Flash 64MB
Network	10/100 Mbps 1Port
Serial Interface	MPC5200 Internal 3 Port, External 4 Channel
USB	USB 2.0 OTC Controller
Storage Device	CF Memory, ATA IDE HDD
Operating Temperature	-30 ~ 75℃
Operating System	Linux

(b-1) Specification of Base-Station H/W

(b-2) Main-Board of Base-Station

(b-3) Case of Base-Station

(b) Specification and H/W appearance of a base-station

Figure 2. S/W and H/W of a base-station

TABLE I. STRUCTURE OF BASE-STATION PROCESSES

Module Name	Description
Sensor Data Acquisition Process (SDAP)	This process receives vehicle detection data from gate-nodes or sensor nodes, and checks validity of the detection. If valid, this process sends them to the shared memory (SM).
Shared Memory (SM)	This is a memory structure to share data among several processes efficiently. Some locks are used to prevent data corruption and to provide synchronized access.
Sensor Data Processing Process (SDPP)	This process extracts more valuable information, for example, speed of the vehicle, from the received detection data. Error information is logged.
Data Transmission Process (DTP)	This process transfers some data and information to external devices such as data management servers in data management layer, or monitoring agents.
Watch-Dog Process (WDP)	This process checks whether all other processes in a base-station operate normally or not. If there are some errors, this process triggers reset procedures for each process.
Base-Station Management Process (BSMP)	This process manages overall operations in a base-station. Target of the management includes operation status, storage availability, RTC (Real Time Clock) status, and so on.

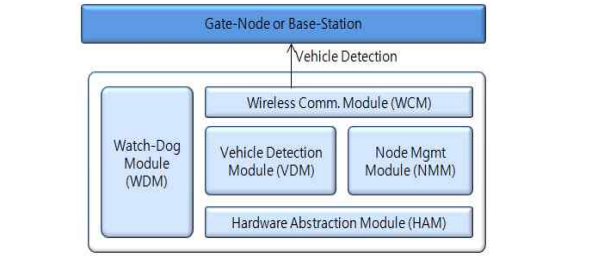
D. *Wireless Sensor Networks*

Wireless sensor network consists of some magnetic sensor-nodes and gate-nodes. The gate-nodes enlarge

TABLE II. STRUCTURE OF SENSOR NODE MODULES

Module Name	Description
Hardware Abstraction Module (HAM)	This module abstracts interfaces for hardware setup and several node operations.
Vehicle Detection Module (VDM)	This module decides whether a vehicle exists or not by executing the pre-defined vehicle detection algorithm. Some filtering mechanisms are used to filter noise.
Node Management Module (NMM)	This module controls execution status of a sensor-node. Also, this module provides recovery procedures and configuration backup.
Watch-Dog Module (WDM)	This module enforces reset options if some runtime errors occurred during node operations.
Wireless Communication Module (WCM)	This module transfers the vehicle detection data to outside of the sensor-nodes. If packet loss in the transmission occurred, it tries to resend by buffering and keeping acknowledgement status. TDMA mechanism [7] with time synchronization protocol is used for packet transmissions.

 ISBN: 978-1-61208-229-5



(a) S/W structure of a sensor-node

Item	Description
MCU	MSP430F2618T
RF Chip	CC2520 [8]
Magnetic Sensor	HMC1043 [9]
Battery	D-size

(b-1) Specification of a Magnetic Sensor Node H/W

(b-2) Appearance of a Magnetic Sensor Node

(b-3) Case of a Magnetic Sensor Node

(b) Specification and H/W appearance of a sensor-node

Figure 3. S/W and H/W of a sensor-node

communicable scope by relay some data packet from sensor-nodes to a base-station. Table II and Figure 3 show the S/W structure and H/W aspects of a sensor-node (The structure and specification of a gate-node is similar to those of a sensor-node except that a gate-node do not include magnetic sensors.) Normally, sensor nodes for vehicle detection are installed on surface of roads, and gate-nodes are installed on a pole on road-side. Therefore, sensor node packaging is different from that of a gate-node. Figure 2(b-3) shows the package of a sensor node.

To provide energy efficiency property, a sensor-node operates on sleep-and-wakeup concept to enlarge life-time of sensor nodes [10]. In other words, when there is no vehicle or specific events, the sensor node enters to sleep mode in which non-essential H/W components are in sleep, and if a vehicle is approaching or some timers are expired, the sensor node wakes up to process the required operations. The figure 4 shows this concept. Moreover, to maximize energy efficiency, more sophisticated method called "dynamic sleep-and-wakeup interval" is used. With the concept, the sensor nodes change some intervals of some node operations dynamically according to existence of a moving vehicle. That is, if it is decided that a moving vehicle exists, the signal sampling operations operates are executed on normal interval, and if it is assumed that there is no a vehicle on the sensor node, the interval of sampling operation is adjusted larger than that of normal.

E. Signal Processing and Vehicle Detection

A sensor-node detects a vehicle by analyzing variations of magnitude of magnetic fields occurred by a moving vehicle. Therefore, a baseline - the level of raw signal which can be obtained when there is no vehicles - is needed to be compared some raw signals acquired when a vehicle is moving. To reduce the complexity of magnetic signal

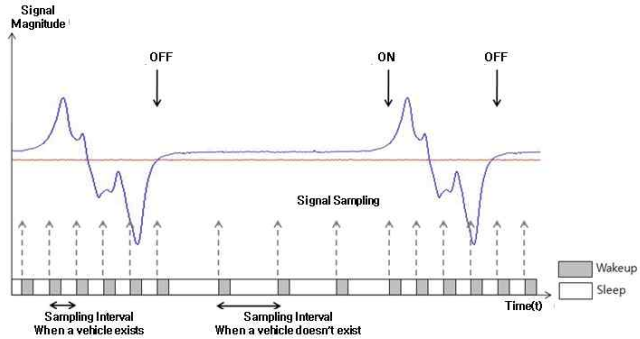
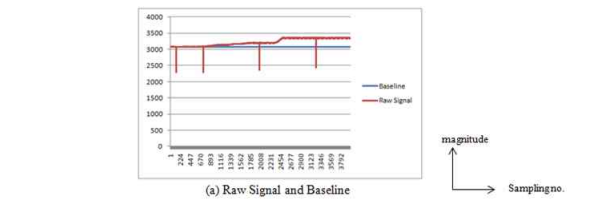


Figure 4. Dynamic sleep-and-wakeup interval mechanism

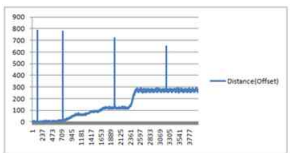
processing and vehicle detection algorithms, not raw magnetic signal values is considered, but pre-processed signals are used for vehicle detection.

Raw magnetic signals include some noises. Such noises should be removed to detect moving vehicles more clearly. There can be various methods to remove noises in raw signals. For example, the calculation of the average of recent N signals is the easiest method to alleviate noise effect. However, the average filter has a disadvantage, that is, it spreads the effect of noises to near signal samplings. In this paper, we adopt a method called interval-based average filter to remove noises. In the filter, if the difference of the magnitude of two signals – the first and the last of a given interval – is smaller than normal noise, all the signals between the two signals modified to the average of the two signals. To remove noises effectively using the interval-based average filter, the interval should be small enough to catch only noises. This filtering method doesn't spread the effect of noises to other signals and the length of the interval (width) and the normal noise level can be algorithm tuning parameters. Figure 5 shows the mentioned signal filtering concepts.

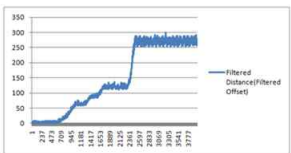
After filtering the raw magnetic signals, the filtered signal becomes to input data of the designed vehicle detection algorithm. In this paper, pattern-based vehicle detection algorithm is designed and used. Increase and decrease pattern of filtered signals can be abstracted as UP and DOWN pattern of which descriptions are followings.



(a) Raw Signal and Baseline



(b) Distance (Offset) between Raw Signal and Baseline



(c) Filtered Distance (Offset) between Raw Signal and Baseline

Figure 5. Magnetic signal filtering and processing for vehicle detection

 ISBN: 978-1-61208-229-5

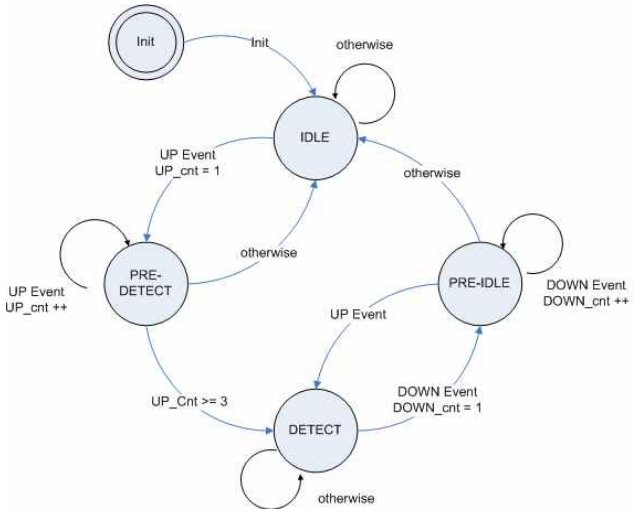


Figure 6. The state transition diagram to detect a moving vehicle

- UP
  The magnitude of magnetic signal increased more than given height during the given successive sampling interval.

- DOWN
  The magnitude of magnetic signal decreased more than given height during the given successive sampling interval.

The vehicle detection algorithm decides whether a vehicle exists or not based on the patterns of UP and DOWN. That is, for example, if successive UPs more than the number of a given value (as a parameter) are given, the algorithm decides that a vehicle is approaching, and, in like manner, if successive DOWNs more than the number of a given value (as a parameter) are given, the algorithm decides that a moving vehicle goes away from the position of the sensor-node. Figure 6 below shows such detection algorithm as a form of state transition machine diagram.

F. Speed Calculation

A base-station calculates speed \using vehicle detection data from sensor-nodes, distance between two sensor-nodes, and time data (clock ticks) of the vehicle detection. However, because the designed vehicle detection system collects data from wireless sensor network using TDMA communication mechanism, there can be several exceptional cases like below. The exceptions should be carefully designed because such exception processing can affect accuracy and performance of overall detection systems and services.

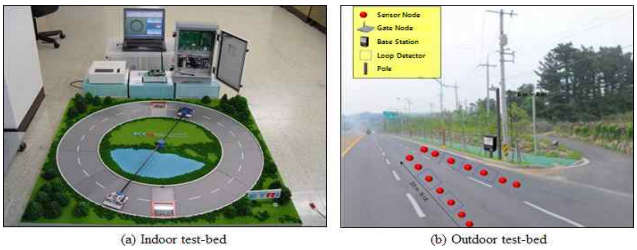
- Time synchronization error between the magnetic sensor network and a base-station
- Loss of ON or OFF data during wireless communication\
- Duplicated transmission of ON or OFF data due to possibilities of packet loss.

III. INDOOR AND OUTDOOR TEST-BEDS

To test and evaluate the vehicle detection system designed and implemented in this paper, several indoor tests and outdoor tests were prepared.

For the indoor tests, a small test-bed was prepared. Several magnetic sensor nodes and a small loop detector were installed on the indoor test-bed, and a model car rotates on the indoor test-bed. The rotation speed of the model car is controlled by external notebook and controller software. Figure 7(a) shows the indoor test-bed. The indoor tests was conducted during about 4 days, the overall accuracy of speed obtained by vehicle detection system related to that of the small loop system was about 98%, although some minor errors were also detected. Because indoor space had clean radio environments and there was just a test car, there were almost no transmission error and no packet loss.

To test the designed system in real-road environments, an outdoor test-bed was constructed as shown at the figure 7(b).



(a) Indoor test-bed (b) Outdoor test-bed

Figure 7. Indoor and Outdoor test-beds

The outdoor test-bed has eight magnetic sensor nodes, one gate-node and one base-station, and one loop system at the same place. The base-station compares the speeds obtained from the designed vehicle detection system and the loop system. The outdoor test is on-going and the results of the tests are being monitoring using a web-site. The overall results will come out about four month later. Analysis on the results of outdoor test-beds will be conducted intensively in various ways, and next paper will cover the analysis results.

IV. CONCLUSIONS AND FUTURE WORKS

In this paper, a vehicle detection system using magnetic sensor network was designed and implemented. Some requirements to be satisfied by the designed system were summarized. The S/W and H/W specification and design were introduced. Also, to test the designed system some indoor and outdoor test-beds were constructed. The result of indoor-tests was excellent, and several outdoor tests are on-going. As future works, some hard test in real-road environments will be conducted. As research topics, real-time vehicle classifications and accurate tracking of a moving vehicle are being considered.

ACKNOWLEDGMENT

This work was supported in part by the IT R&D program of MIC/IITA [10035249, Development of U-TSD (Traffic Surveillance & Detection) Technology] which are Korea government department regarding to Information and Communication.

 ISBN: 978-1-61208-229-5

REFERENCES

[1] Kanian, A.N., "A wireless sensor network for smart roadbeds and intelligent transportation systems," M.Eng. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, June 2000.

[2] Marcin K., Aline S., and Vinny C., “Sensor Networks for Smart Roads,” PerCom 2006 Workshop, Pisa, Italy, March. 13-17. pp. 306-310, 2006.

[3] Sang K. L. and YoungJun M., “Intelligent All-Way Stop Control System at Unsignalized Intersections,” In Proceeding of ICCME 2004, Greece, Nov. 19-23, pp. 13-18, 2004.

[4] Kim, J. S., Lim. J. H., Pelczar, C., and Jang. B. T., “Sensor Networks for Traffic Safety,” In Proceeding of VTC 2008, Singapore, May. 11-14, pp. 3052-3056, 2008.

[5] Chueng, S. Y., Coleri, S., Dundar, B., Ganesh, S., Tan, C. W., and Varaiya, P., "Traffic Measurement and Vehicle Classification with a Single Magnetic Sensor," Journal of Transportation Research Board, pp. 173-181, Feb., 2005.

[6] Yuhe Z., Xi H., Li. C., and Ze Z., “Design and Evaluation of a Wireless Sensor Network for Monitoring Traffic,” In Proceeding of ITS World Congress, Beijing, China, Oct. 9-13, pp. 1258-1263,2007.

[7] Saurabh. G., Ram. K., and Mani. B. S., “Timing-sync Protocol for Sensor Networks,” In Proceeding of SenSys 2003. LA, USA., Nov. 5-7, pp. 138-149, 2003

[8] http://www.ti.com/product/cc2520 [retrieved: Oct. 2012]

[9] http://www51.honeywell.com/aero/common/documents/myaerospacecatalog-documents/Missiles-Munitions/HMC_1001-1002-1021-1022_Data_Sheet.pdf [retrieved: Oct. 2012]

[10] Joseph, P., Robert, S., and David, C. "Telos: Enabling Ultra-Low Power Wireless Research," In Proceedings of 4th IPSN, pp. 364-369, 2005.

 ISBN: 978-1-61208-229-5

Interference-aware Supermodular Game for Power Control in Cognitive Radio Networks

Sonia Fourati, Soumaya Hamouda, Sami Tabbane
Lab. MEDIATRON
Communication Engineering School (Sup'Com)
Tunis, Tunisia
sonia.fourati@yahoo.fr,
{soumaya.hamouda, sami.tabbane }@supcom.rnu.tn

Miquel Payaro
Centro Technologico de Telecommunicacions de Catalunya (CTTC)
Spain
miquel.payaro@cttc.es

***Abstract*—This Cognitive radio has proved to be a promising solution to improve the utilization of the radio spectrum. This new concept allows different wireless networks to operate on the same spectrum bandwidth. An efficient power control is thus crucial to make this coexistence possible and beneficial. In fact, unlicensed users (or secondary users, SUs) should communicate without harming the Primary users' (PUs) transmissions. In this paper, we propose a new power control mechanism for the SUs based on a powerful mathematical tool, the Game Theory. Our algorithm is based on a non-cooperative supermodular power control game in which we define a new utility function under total transmit power constraint, but also under interference constraint. We prove the existence of the Nash Equilibrium analytically and by means of simulations.**

Keywords- cognitive radio networks; game theory; power allocation; supermodular game.

I. INTRODUCTION

The wireless networking technologies are evolving rapidly in a very diverse manner (e.g., 3G+ [2] and 4G [3] cellular networks). This dramatic increase of the demand for spectral bandwidth and quality service is limited by the scarcity of spectrum resources. The CR (Cognitive radio) [4] is viewed as an effective approach for improving the utilization of the radio spectrum. The main idea of CR is to make secondary users (SUs), equipped with smart radios, able to sense the environment, detect the unused spectrum resources (spectrum holes) and decide when and how to access these holes. The CR also permits to the SUs to have underlay access to spectrums at the same time with the primary user, without causing harmful interference to the latter. Therefore, new challenges related to spectrum sharing appear especially how to design efficient power allocation and channel assignment schemes.

Many researches were conducted to resolve the topic of power control for Cognitive Radio Networks (CRNs) with the powerful mathematical tool the Game Theory [1]. In fact, Game Theory, having proved its efficiency in economics, was introduced lately to solve problems related to radio resource allocation in telecommunications [1]. This instrument helps study the complex interactions among independent players in order to optimize the setting of various elements of the network. More precisely, the players can be the SUs in a CRN, and the various elements are the SUs' transmit power. Game Theory will thus help us to design and to resolve the topic of power control for underlay SUs' transmission in presence of PUs under interference level constraint. It will also allow us to investigate the existence and convergence to a steady state operating point called Nash Equilibrium (NE), when the SUs perform independent distributed adaptations in terms of transmit power.

Some power control games are non-cooperative games, such [5] and [6]. In these games, selfish users choose their transmit power and attempt to maximize their individual utilities without being careful about the impact of their strategies on other users. A typical solution to a non-cooperative game is the Nash Equilibrium Point (NEP) [1], which is an equilibrium point where each player has no chance to increase its utility by unilaterally deviating from this equilibrium. Saraydar et al. [5] propose a non-cooperative power control game model in CDMA networks and proves the existence and uniqueness of Nash equilibrium. Zhou et al. [6] considered the problem of joint power and rate control for SUs in cognitive radio network by using non-cooperative game theory given a certain QoS requirement of SUs. Rasti et al. [7] proposed a non-cooperative power game with pricing that is linearly proportional to the signal-to-interference ratio. Del Re et al. [8] present a power resource allocation technique based on game theory, considering mainly Potential Games. Such power allocation is aimed to the up-link communication in a centralized CRN. Jing and Zheng [9] presented a game theoretic solution for uplink resource allocation in multi-cell OFDMA systems. Steady state and convergence are analyzed with potential game. The game can be modeled as a potential game to guarantee the convergence of NE. Del Re et al. [10] provided an S-Modular game in order to solve the resource sharing between the PU and the SU in a distributed and fair way. Elias et al. [11] use the Game Theory to address the spectrum access for the SUs taking into account the congestion level observed on the available spectrum bands. The same authors also proposed [12] a joint pricing and network selection scheme in CRNs based on a Stackelberg (leaderfollower) game.

In this paper, we are interested in developing a new algorithm based on Game Theory for a distributed power

 ISBN: 978-1-61208-229-5

control in CRNs. This algorithm should allow the SUs to transmit without harming the PUs communications and while guaranteeing SU's Quality of Service requirements. We use a supermodular game because the latter guarantees the existence of at least one NE to reach, according to the different best SUs' responses.

This paper is organized as follows: Section II introduces the game model including the system model and formulates the optimization problem. In Section III, we describe our distributed power allocation algorithm based on our proposed supermodular game theoretic. Simulation results are given and discussed in Section IV. Finally, Section V concludes this paper.

II. GAME MODEL

Game theory analyzes the strategic interactions among rational decision makers. Three important components in a game model are the set of players, the strategy space of each player, and the payoff/utility function, which measures the outcome of the game for each player.

As the cognitive radios are smart terminals that can learn from their environment and dynamically modifying their transmission parameters in order to optimize their performance. Therefore, their interactions can be modeled using a non cooperative power control game. In this game model, $G = \left\{ N, \{P_i\}_{i \in N}, \{U_i\}_{i \in N} \right\}$, the N secondary users are the players and $P_i = \left\{ p_i, 0 \le p_i \le P^{\max} \right\}$ are their strategies, which represents the sets of power allocation that influence their own performance. P^{max} is the SU maximum power. And, U_i is the desired performance, designed as the payoff or utility function.

A. System Model

We consider a cognitive radio system in which a primary network is consisting of one PU base station and M active PUs coexists on one hand, and a secondary cognitive network made by N SUs equipped with CRs in a spectrum underlay manner on the other hand. SUs can simultaneously transmit with PUs but have to strictly control their transmit power to avoid harmful interference with PUs. We denote each transmitter and its intended receiver pair by a single index i (i = 1,…., M), referred to as a user. For simplicity, we neglect the interference from other adjacent cells. Thus, for each pair of SUs (v_i, v_j) located within mutual communication range, the signal-to-interference-and-noise ratio (SINR) received at user i can be written as:

$$\gamma_i = \frac{h_{ii} p_i}{\sum_{j=1, j \neq i}^{N} h_{ji} p_j + \sum_{m=1}^{M} h_{mi} p_m + \sigma^2} \quad (1)$$

where h_{ii} and p_i are respectively the channel gain, and the power level for the i^{th} player SU, in watts and is a parameter that we used in this paper for power control between [0, p^{max}]. h_{ji} is the cross channel gain from transmitter j to receiver i. The channel gain is determined by the log-normal shadowing path loss model. P_j is the transmission power of other SU different from SU i. and p_m is the transmission power of PUs and σ^2 is the additive white Gaussian Noise power (watts). Then, the transmission rate of the SU i at time t is:

$$R_i(t) = W \log_2 \left[1 + \eta \gamma_i \right] \quad (2)$$

where η is the SNR gap and it is related to the BER, bit error rate. It is given by $\eta = -1.5 / \ln(5*\text{BER})$[12].

B. Utility Function Design

In this section, we seek to design a proper utility function that not only reflects the benefit of the player but also facilitates the implementation of power control algorithms in terms of convexity and global convergence. The key is to find utility expressions that are not only physically meaningful for a CRN but mathematically attractive for ensuring global convergence to the NEP [13] as well. In this paper, we adopt the $R(t)$ in eq.(2) as the QoS metric for SU players and accordingly construct the novel utility function U_i as an SINR-related form. The utility function represents the future benefit that a player will achieve when adopting a certain strategy, i.e., power allocation. However, the overall network optimum is usually not achieved at the NEP, since selfish users are only interested in the individual benefit. To improve the efficiency of the NE of non-cooperative games in CRNs, pricing can be introduced when designing the non-cooperative game, in order to guide the selfish users to a more efficient NE [14].

Each SU maximizes its own data rate at the cost of high power consumption, which causes interference to other SUs and brings down their data rate. In order to keep a SU from selfishly transmitting the highest transmit power, the system should first impose certain throughput fairness among the SUs, but also a pricing function. In our paper, we propose that this pricing function reflects constraints on the SU transmit power as well as constraints on the interference level caused on a PU. The utility function is therefore expressed by:

 ISBN: 978-1-61208-229-5

$$\text{(P1):}\quad u_i\left(p_i, p_{-i}\right) = N\log\left(R_i\right) - \beta p_i - \alpha h_{im} p_i \qquad (3)$$

$$s.t\ \gamma_i \geq \gamma_{\min} \qquad (a)$$

$$\sum_{i=1}^{N} h_{im} p_i \leq I_{th}, \quad m = 1,....,M \qquad (b)$$

$$0 \leq p_i \leq P^{\max}, \quad i = 1,....,N \qquad (c)$$

where R_i is the achieved SU throughput, β is a positive constant, considered as the price of each SU transmit power p_i. The second part of (3), $\alpha h_{im} p_i$, considers the interference caused on the PU m by the user SU i. The constraint (a) reflects the minimum required quality of service for each SU; and the constraint (b) reflects that the aggregated interference caused at the PU should be below a predefined threshold I_{th}.

On the other hand, the power allocation can be formulated as an optimization power control problem given by:

$$\max_{p_i} u_i\left(p_i, p_{-i}\right) \qquad (4)$$

$$s.t. \quad \gamma_i \geq \gamma_{\min}$$

$$\sum_{i=1}^{N} h_{mi} p_i \leq I_{th}, \quad p = 1,....,M$$

$$0 \leq p_i \leq P^{\max}, \quad i = 1,...,N$$

C. Existence of Nash Equilibrium

The NE gives the best strategy given that all the other players stick to their equilibrium strategy too. However, the question is how to find the Nash equilibrium, especially when the system is implemented in a distributed manner. One approach is to let players adjust their strategies iteratively based on accumulated observations as the game unfolds, and hopefully the process could converge to some equilibrium point.

The NE is the steady state in the game, in which no player can increase its utility function from unilaterally deviating its action. However, it does not follow that there is a NE in every game. Therefore, it becomes necessary to prove the existence of NE. For example, when the game can be modeled as a super modular game, convergence to the NE is guaranteed.

Theorem:

Our proposed game model can be shown as a supermodular game.

Proof: 1) Since [0, P^{max}] is a compact subset of R, 2) Also, for the range of 0≤ P_i ≤P^{max}, the utility function is continuous. 3) In addition, the utility function chosen has an attractive property: it is twice differentiable. The remaining condition that we should check is whether $\partial^2 U(Pi)/\partial Pi \partial Pj > 0$ or not.

Let $B = \sum_{\substack{j=1 \\ j \neq i}}^{N} h_{ij} p_j + N_0$ then the partial differential form of the above payoff function is:

$$\frac{\partial U(p_i)}{\partial p_i} = \frac{N h_{ii}}{\left(B + h_{ii} p_i\right)\log\left(1 + \frac{h_{ii} p_i}{B}\right)} - \beta - \alpha h_{im} \qquad (5)$$

Let $C = \left(B + h_{ii} p_i\right)\log\left(1 + \frac{h_{ii} p_i}{B}\right)$, then:

$$\frac{\partial^2 U(p_i)}{\partial p_i \partial p_j} = \frac{-N h_{ii} h_{ij}\left(\log\left(1 + \frac{h_{ii} p_i}{B}\right) + \frac{h_{ii} p_i}{B}\right)}{B^2} \qquad (6)$$

Since $\log(1+x) < x$ for all $x>0$, and $(h_{ii}p_i/B)$ is positive; therefore, $\frac{\partial^2 U(p_i)}{\partial p_i \partial p_j} > 0$. According to the definition and property of game modes, this game is a supermodular game and therefore must be at least one NE in this supermodular game.

D. Solution of the game

Since the existence of NE was proved, we consider the problem of how to identify it. The optimal transmit power or NE can be obtained in such a way that each SU maximizes its own utility function iteratively. The problem can be expressed s follow:

$$P_i^* = \arg\max U_i\left(p_i, p_{-i}\right) \quad , i \in \mathrm{N} \qquad (7)$$

where $P_i^* \in \left[0, P^{\max}\right]$

It should be noted that there is no sufficient guarantee in this game with regard to constraint (a), (b) and (c) of the problem (P1). First, the protection of PU should be assured by keeping the interference below a threshold, and the rigid SINR requirement of each SU must be respected especially if the SU experiences strong interference. In the next section, we will give details to solve this problem.

III. DISTRIBUTED POWER ALLOCATION GAME

In this section, an algorithm based on Lagrange techniques is developed to solve (4). This algorithm will have provable convergence and is suitable for distributed implementation. Because the model relates the optimum solution with 3 constraint conditions, let λ_i and μ_i denote

 ISBN: 978-1-61208-229-5

Lagrange multipliers corresponding to minimum SINR constraints (a) and the interference constraints (b) respectively.

The Lagrangian function of the convex equivalent of (3) is then:

$$L(p,\lambda_i,\mu_i)=U_i(p_i)+\sum_{i=1}^{N}\lambda_i(\gamma_i-\gamma_{\min})+\sum_{i=1}^{N}\mu_i\left(\sum_{i=1}^{N}h_{mi}p_i-I_{th}\right) \quad (8)$$

The problem (P1) is equivalent to:

$$\max_{p} L(p,\lambda_i*,\mu_i*)$$
$$s.t \quad 0\le p_i\le P^{\max}, i=1,2,\ldots,N \quad (9)$$

The problem (P2) is solved via the following first-order algorithm that utilizes the gradient of L(p,λ,μ) to simultaneously update primal and dual variables with constant step size β and [x]+=max{0,x} :

$$p_i(k+1)=p_i(k)+\beta\frac{\partial L(p,\lambda,\mu)}{\partial p_i} \quad (10)$$

$$\lambda_i(k+1)=\left[\lambda_i(k)+\beta\gamma_i\right]^{+}$$
$$\mu_i(k+1)=\left[\mu_1(k)+\beta\sum_{i=1}^{N}h_{mi}p_i\right]^{+} \quad (11)$$

The gradient $\nabla L(p,\lambda,\mu_0,\mu)$ is used in (8) to find the maximum of $\nabla L(p,\lambda,\mu_0,\mu)$ with respect to p, and convergence will lead to the NE.

IV. PERFORMANCES EVALUATION

To evaluate the performances of the proposed algorithm, the simulations have been performed with a reduced number of users. Just one PU receiver has been placed in the scenario, while at most five SUs have been considered for the secondary system. The cell radius is R = 500m. The propagation model takes into consideration of path loss and frequency selective fading. The background noise $\delta 2$ is 5×10^{-15} Watts. The transmit power of PU is 10Watts .In such a scenario, the game converges quickly to the Nash equilibrium after 2-3 iterations.

First, we examine the convergence performance of the proposed game model in terms of SU transmit power. Fig. 1 illustrates the evolution of the SU transmit power for the five secondary users. It shows that the transmit power for each SU converges to the steady state. From this figure, we observe that there all the five SUs are transmitting with reasonable at the maximum power. This can not only enhance the power consumption for these SUs but also reduced the level of the interference to the PU. The limitation of the overall interference in the system is thus achieved.

Fig. 2 illustrates the achieved throughput by the different SUs versus their quality of service requirement in terms of BER. In fact, these SUs have not only satisfied the quality of service requirements but also realized a total throughput in the system of almost 14 Mbps.

TABLE I. THE LIST OF PARAMETERS FOR A SINGLE CELL COGNITIVE SYSTEM

Parameters	***Value***
W, the spectrum bandwidth	5 Mhz
Cell Radius	500 m
Number of users	5
Pmax, maximum power constraint	1 Watt

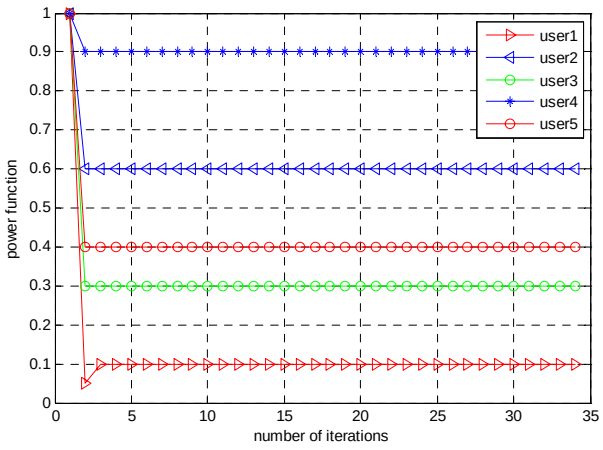


Figure 1. Convergence of SUs' transmit power.

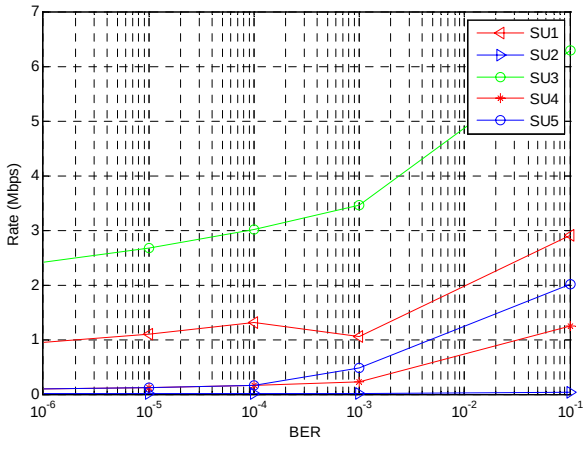


Figure 2. Achieved SUs' throughputs vs. SUs' target BERs

V. CONCLUSION AND FUTURE WORK

In this paper, a non cooperative power control game is investigated for CR networks under quality of service and interference constraints. More precisely, we have introduced a new utility function in which the constraint on the interference caused by the SU to the PU is considered as well as the SU transmit power limitation. We have proved the NE for our game, and gave a distributed power control algorithm that converges to the NE.

The proposed algorithm used a pricing-based game to achieve the efficient power control which resulted in the maximum throughput for the cognitive network and respected the interference limitation as well. In the future, we intend to efficiently modify the price function so that we could maximize the throughput without altering the transmit power. Also, we intend to maximize the overall system throughput using cooperation between SUs.

REFERENCES

[1] D. Fudenberg and D. K. Levine, "Game theory", MIT Press, Cambridge, MA, 1991.

[2] http://www.3gpp.org

[3] ITU-R Circular Letter 5/LCCE/2, "Invitation for submission of proposals for candidate radio interface technologies for the terrestrial components of the radio interface(s) for IMTAdvanced and invitation to participate in their subsequent evaluation," March 2008.

[4] J. Mitola III, "Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio," Doctor of Technology Dissertation, Royal Institute of Technology (KTH), Sweden, May, 2000

[5] U. Cem Saraydar, B. NarayanMandayam, and J. David Goodman "Efficient power control via pricing in wireless data networks," IEEE Transactions on Communications, vol. 50 (2), pp. 291 – 303, 2002

[6] P. Zhou, W. Yuan, W. Liu, and W. Cheng, "Joint power and rate control in cognitive radio networks: A game-theoretical approach," in Proc. IEEE Int. Conf. Commun., pp. 3296–3301, 2008.

[7] M. Rasti, A. R. Sharafat, and B. Seyfe, "Pareto-efficient and goal-driven power control in wireless networks: A game-theoretic approach with a novel pricing scheme," IEEE/ACM Trans. Netw., vol. 17, no. 2, pp. 556– 569, Apr. 2009.

[8] E. Del Re, G. Gorni, L. Ronga, and R. Suffritti, "A power allocation strategy using Game Theory in Cognitive Radio networks," in Proceedings of the First ICST international conference on Game Theory for Networks, GameNets'09, pp. 117 - 123 , 13-15 May 2009.

[9] Q. Jing and Z. Zheng, "Distributed Resource Allocation Based on Game Theory in Multi-cell OFDMA Systems", International Journal of Wireless Information Networks, Vol. 16, pp.44-50. March 2009.

[10] E. Del Re, R. Pucci and L. S. Ronga, "Energy Efficient Non-Cooperative Methods for Resource Allocation in Cognitive Radio Networks" , communications and network journal, February 1, 2012

[11] J. Elias, F. Martignon, A. Capone, and E. Altman, "Non-Cooperative Spectrum Access in Cognitive Radio Networks: a Game Theoretical Model," Computer Networks, Elsevier , vol. 55, issue 17, pp. 3832-3846, December 2011.

[12] J. Elias, F. Martignon, and E. Altman, Joint Pricing and Cognitive RadioNetwork Selection: a Game Theoretical Approach, accepted forpublication in WiOpt 2012, Paderborn, Germany, May 2012.

[13] C. Yang, J.D. Li, and Zhi Tian, "Optimal Power Control for Cognitive Radio Networks Under Coupled Interference Constraints: A Cooperative Game-Theoretic Perspective", IEEE Transactions on Vehicular Technolgy, Vol. 59, No. 4, May 2010

[14] B. Wang , Y. Wu, and K.J. Ray Liu," Game theory for cognitive radio networks: An overview", Computer Networks , Volume: 54, Issue: 14, pp. 2537-2561 , 2010.

 ISBN: 978-1-61208-229-5

Wireless Communications Enabling Smart Mobility: Results from the Project PEGASUS

Alessandro Bazzi, Barbara M. Masini, Gianni Pasolini, and Oreste Andrisano
CNR-IEIIT and DEIS, University of Bologna @WiLab
Bologna, Italy
Email: {alessandro.bazzi, barbara.masini, gianni.pasolini oreste.andrisano}@unibo.it

***Abstract*—Wireless communications for real time traffic information are considered to efficiently provide smart mobility in congested cities. In this work, we aim at summarizing objectives the results of the Italian project PEGASUS, where wireless communications have been exploited, in real time, to: i) acquire traffic information directly from vehicles (uplink) and ii) retransmit updated information to interested vehicles (downlink) after a proper processing at a control center. Specifically we focus on i) the uplink collection of data from vehicles through the universal mobile telecommunication system (UMTS), ii) the downlink transmission of updated information to vehicles through UMTS, iii) the cellular resource saving through the exploitation of short range communications based on wireless access in vehicular environment (WAVE)/IEEE 802.11p, and iv) the impact of updated information on travel time. Results are provided through the development of an integrated simulation platform that jointly takes into account vehicular traffic behavior in urban environment, data processing at the control center, and performance of the communication networks at the different layers of the protocol pillar.**

***Keywords*-Intelligent transportation systems (ITS); real time services; simulations in realistic scenarios.**

I. INTRODUCTION

Keeping traffic moving is a challenge that governments, industries and researchers are facing worldwide nowadays. Effective solutions can only be obtained with a capillary real time knowledge of the traffic conditions and a prompt communication to the drivers; without updated and dynamic traffic information, only particular events or repetitive situations can be handled. The creation of an infrastructure for communication between vehicles, service centers and sensors, is thus one of the main needs identified by international institutions, service providers and car manufacturers to address with satisfactory results the problems generated by traffic, justifying the big efforts that are being pushed both in Europe and in the rest of the world [1]. To this scope, different wireless access technologies could be exploited, from short-range ad-hoc networks to cellular systems. Regarding the former ones, wireless access in vehicular environment (WAVE) [2], based on IEEE 802.11p [3] represents the future for vehicle-to-vehicle (V2V) communications. This technology, well suited for safety applications, entertainment, gateway access, and road charging, can also be used for traffic management service, on condition that a connection to a remote control center is available. Vehicles, in fact, must periodically collect their position and speed and send such data to a control center, that is in charge of retransmitting back aggregated traffic information. Even in the case that all vehicles were equipped with such technologies, the need for an infrastructure makes the adoption of IEEE 802.11p for traffic management services a long term solution, due to the investment that the deployment of a communication infrastructure requires.

Hence, thinking to short term, cellular systems appear as the only feasible solution, already guaranteeing high penetration and wide coverage worldwide, also allowing continuity of service at vehicular speeds. And this is particularly true noting that, on the one hand, the last generation of on board navigators are already equipped with a cellular interface, and, on the other hand, smart phones embed navigation functionalities (often for free). However, the expected increase of vehicles equipped with on board units (OBUs) could lead to an overload of the cellular access network, and, consequently, to a degradation of the quality of service provided to voice and data users [4].

This work is carried out in the framework of the Italian project PEGASUS [5] and aims at summarizing the obtained results. Most of them have been already published (see, e.g., [4], [6], [7], [8], [9], [10]), but this is the first paper that summarizes the project as a whole. PEGASUS relies on over one million vehicles already equipped with OBU periodically transmitting their position and speed to a control center. Considering this scenario, we focus on i) the uplink transmission of data from vehicles through cellular systems, ii) the downlink transmission of updated traffic information to vehicles through cellular systems, iii) the exploitation of short range V2V and vehicle-to-roadside (V2R) communications to save cellular resources, and iv) the impact of updated traffic information on travel time.

In particular, firstly focusing on the universal mobile telecommunications system (UMTS) as the enabling technology for uplink and downlink information, we aim at:

- investigating the feasibility of the acquisition of small but frequent amount of data from many vehicles (*uplink performance*). Is the UMTS capacity sufficient for these kinds of multiple connections?

 ISBN: 978-1-61208-229-5

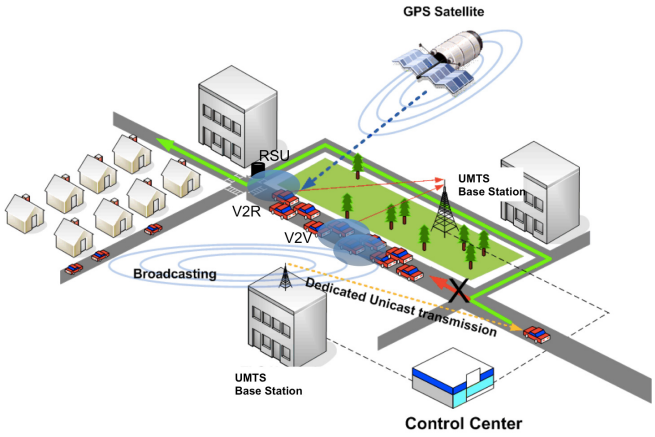


Figure 1. Scenario: real time traffic information exchange and enabling technologies.

- verifying the feasibility of real-time transmissions to vehicles of traffic information, both in unicast and multicast mode (*downlink performance*);
- investigating the impact of this service on the others already supported by the network (both in uplink and downlink). Is the quality of service (QoS) perceived by the final user sufficient, or does the network (or the service) need enhancements?
- evaluating the QoS perceived by the users of the traffic management service.

Then, considering also WAVE/IEEE802.11p-based V2V and V2R communications, we:

- investigate the impact of V2V and V2R communications (*short range communications*) to share and aggregate traffic information with the aim to reduce the cellular network load and the delivery delay. How many road side units (RSUs) are needed? which is the impact of multi hop V2V communications?

Finally, to also evaluate the impact of real time communications on traffic management, we:

- estimate the travel time of vehicles equipped with smart navigators (*smart navigation service*).

To address these issues, we developed a realistic simulation platform taking into account the details of the protocol pillar, the mobile propagation conditions, and the users' mobility.

The paper is organized as follows: In Section II related works are reported; In Section III, the considered scenario is depicted and the main characteristics and motivations of enabling technologies are provided; In Section IV, the simulation platform developed to investigate the considered scenario is described; In Section V, the simulation settings and output figures are given; In Section VI, numerical results are provided; Finally, in Section VII conclusions are drawn.

II. Related Works

Cellular systems are nowadays widely recognized as drivers of innovation in a wide range of technical fields, and represent the shortest term solution to collect data from vehicles and retransmit them to on board navigators avoiding new set-ups or expensive installations [11], [12]. Various activities are ongoing [6], [7], [13], [14], and products based on cellular technologies are already on the market.

A. Uplink

Some studies on the cellular performances in vehicular applications are coming out (see, e.g., [11]), but still few investigations are performed on the impact that these new services have on other cellular services (such as voice) in terms of resource sharing and consequent QoS guaranteeing. A study on the feasibility of data acquisition from vehicles through cellular systems has been performed in the German project Aktiv CoCar [15], [16], that defined a new protocol, called "traffic probe data protocol" (TPDP), to upload traffic data through UMTS common channels. However, results are given only in terms of cumulative distribution function (CDF) of the end-to-end delay, and no evaluation of the impact of this service on the network and user's QoS is given, which is, in turn, provided in [4] and here summarized and reported in the context of the PEGASUS project.

B. Downlink

Focusing on communications from a control center to vehicles, due to the widespread diffusion of connected on-vehicle navigators and smart phones with positioning applications, cellular systems represent the short term solution to solve the problem of real time traffic information to and from vehicles, so to suggest alternative routes and avoid congestions. This is confirmed by recent publications [11], [17], [18] and market trends: many vehicles are worldwide equipped with connected smart navigators receiving via general packet radio service (GPRS) updated traffic information. Again focusing on UMTS, the objective is to investigate which solution between unicast or multicast transmission has to be preferred and which is the impact on the QoS experienced by vehicular users and other users.

C. Short Range Communications

In the context of V2V and V2R communications, several standardization processes and research works are currently carried out (see, e.g.,[19], [20]) giving particular attention to WAVE [2] based on IEEE 802.11p [3]. In [21], the performance of WAVE/IEEE 802.11p vehicular ad hoc networks (VANETs) are evaluated in terms of packet delivery rate and delay assuming roadside access points. [4], [7], [13] investigate the real time acquisition of traffic information through cellular systems and its impact on both the network load and users' satisfaction when no V2V communications are considered. The novelty in this work is to share and

 ISBN: 978-1-61208-229-5

aggregate traffic information with the aim to reduce the cellular network load and the delivery delay when real time traffic information systems are concerned through the exploitation of short range V2V and V2R communications.

III. Scenario and Enabling Technologies

The considered scenario is shown in Fig. 1: vehicles are equipped with on board units (OBUs) acting as traffic sensors and periodically transmitting their position and speed to a remote control center through the cellular network. The data collected at the control center are processed in real time to evaluate the actual traffic conditions (i.e., the travel time) on each monitored road segment; when the control center detects traffic conditions different from those foreseen by the static roadmap data base, it updates the roads status on a dynamic data base that will be queried for the best route evaluation. Updated measurements are then transmitted from the control center to the interested vehicles through the cellular network. We assume that updated traffic conditions can be transmitted from the control center to on board navigation systems or to the user's personal smart phone, hereafter simply denoted as *smart navigator*. Hence, the smart navigator receives information on the traffic conditions for each road segment from its current position to the destination and calculates the optimal route; on a real time basis, it updates its data and, in case, modifies the route in order to avoid any slowdown.

A. Uplink

Among the cellular technologies, GPRS is nowadays the most adopted for uplink measurements transmission. However, to transmit data over the GPRS network, the mobile station (MS) must first send a message on a common channel asking for a dedicated resource, with procedures requiring a not negligible access time, in the order of seconds [22]; for this reason, OBU collects tens of measurements before transmitting them in a single packet. This approach obviously increases the data acquisition delay at the control center. Differently to this, UMTS also allows the transmission of small amount of data over the shared signalling channel random access channel (RACH), avoiding the set-up of dedicated resources [23]. This way, any measurement can be transmitted by the OBU as soon as it is taken, with minimum delay and reduced signalling overhead. This solution appears promising especially considering the forecasted increase in the number of equipped vehicles, but it clearly requires investigations on feasibility and resources occupation.

Here, we discuss the impact of the real time data acquisition on capacity and coverage of existing cellular systems, foreseeing the realistic perspective of an explosion in the number of equipped vehicles.

B. Downlink

We exploit UMTS as the enabling technology either in unicast mode via dedicated unicast channels (DCHs), or in multicast mode via multimedia broadcast multicast service (MBMS). The objective is to evaluate if the network can support the additional new load and the impact it has on the performance perceived by other UMTS users.

Due to the adoption of code division multiple access (CDMA) [4], the number of active channels in UMTS is a consequence of the trade-off between coverage and capacity, and the amount of resources occupied by each transmission is given in terms of used power: on the one hand, a higher data rate as well as a higher distance from the base requires a higher power for a sufficient QoS; on the other hand, a higher power reduces the cell capacity. The power is, in fact, a limited resource at the base station (in downlink) and each transmission turns into an interference to all other active communications (in both directions).

As far as MBMS is concerned, it allows to share resources among many user. Hence, power is allocated to MBMS channels only once for any number of users in the cell receiving the service. We assume that vehicles are equipped with MBMS units joining the multicast group where traffic-related messages are distributed. It has to be remarked that MBMS uses part of the available power at the base station, thus limiting the number of DCHs that can be established. Moreover, the broadcast/multicast nature of the channel does not allow to exploit the fast power control feature that is of main importance for an interference limited system like UMTS; the base station pre-assigns a certain amount of power to MBMS services depending on the coverage planning and the desired bit rate.

The following strategies, thoroughly described in [7], are assumed for traffic updates:

- For the unicast mode, the update involves *road segments encompassed by an ellipse* whose focuses are the actual vehicle position and either the next intermediate point or the final destination. This strategy avoids the transmission of information related to road segments too far from the actual vehicle position, which would be out of date when the vehicle needs it. Moreover, since only the transmission of the coordinates of two points is needed from the navigation system to the control center, the amount of data transmitted in the uplink is very small, thus limiting costs and resource occupation. Following [7], 1000 road segments are updated every 5 minutes.
- For the multicast mode, a *progressive coverage* strategy is considered, consisting in the transmission to the on-board navigator of the information related to the most important roads at national level and regional level, and to the minor roads at local level only. Following [7], 12000 road segments are updated in average.

Independently on the unicast or multicast communication technology, we assume the adoption of the transport protocol experts group (TPEG) technology [24] at the highest layers of the protocol pillar with 60 bytes packet per each road segment (one packet per direction) [14].

 ISBN: 978-1-61208-229-5

C. Short Range Communications

When UMTS is the only available communication interface, each OBU autonomously transmits collected data through the cellular system either after a given time out or when a certain amount of data has been collected. If also V2V and V2R communications can be exploited, information can be exchanged also between vehicles and between vehicles and RSUs. In this case, OBUs communicate one with each other exchanging information and aggregating redundant data referring to the same road segment. Both the aggregation of information and the transmission through RSUs allow to reduce the UMTS load toward the control center, thus saving the limited capacity of the cellular network and the related costs. We assume that, when vehicles are also equipped with technologies for short range communications, WAVE/IEEE 802.11p [3] is adopted as the standard for communication and OBUs know the positions (i.e., the coordinates) of RSUs and are able to associate any measured data with the correspondent road segment. In particular, the following routing strategy is assumed throughout the paper: communications occur between two vehicles a time; each vehicle identifies among its neighbors the one nearest to a RSU, which is elected as master. The master is in charge of receiving the data, merging the received data with its own if redundant, and transmitting them as far as it is within the coverage distance of the RSU. Since this procedure is performed iteratively, data can rapidly reach a RSU even from a relatively far distance if a sufficient density of vehicles is equipped with OBUs. Packets that cannot be transmitted to a RSU are transmitted via UMTS when one of the following conditions occurs: after a maximum number of packets is reached in the transmission queue or at a given time out. To detect which vehicle is the nearest to a RSU, GPS coordinates are also exchanged.

In the further, we show the advantages achievable by sharing and aggregating the collected traffic information through the exploitation of V2V and V2R communications, and the impact of number and positions of the RSUs.

IV. Investigation Tools

The investigation of the considered scenario requires a complete simulation of the cellular network, both in the uplink and in the downlink, as well as a realistic simulation of vehicles' movements. In fact, the vehicular mobility significantly impacts on the performance of the telecommunication network and on the traffic redistribution itself. A realistic mobility model is thus needed, and it has to take into account all roads, with their speed limits, vehicles acceleration and decelerations, queues at traffic lights, etc. Hence, we developed a simulation platform which integrates VISSIM [25] as vehicular traffic simulator and the simulation platform for heterogeneous interworking networks (SHINE) [26] as cellular network simulator, thus allowing us to provide realistic results both in terms of vehicular traffic

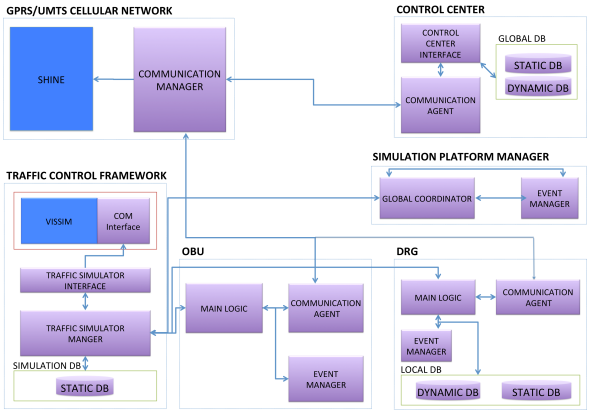


Figure 2. Integrated platform for the simulation of smart navigation.

(with queues, number of lanes, one way roads, etc.) and communication systems [8].

Vehicular simulator: VISSIM [25]. It is a microscopic simulation tool modelling traffic flow in urban areas as well as interurban motorways, and allowing to reproduce car-following and lane changing as in real scenarios. It uses a psycho-physical car following model for longitudinal vehicles movement and a rule-based algorithm for lateral movements. VISSIM allows to be controlled by external applications with the use of a component object model (COM): by the adoption of dynamic link libraries (DLL), it is possible for an application to control the movement of vehicles and to manage the whole simulation.

Cellular network simulator: SHINE [26]. It is an event driven dynamic simulator which allows to jointly take into account the whole cellular network architecture from the application layer to the physical layer, also carefully reproducing the time and frequency correlated behavior of the wireless medium, which is often approximated or even neglected in most network simulator. All relevant aspects of the real system are reproduced, including for example a not uniform positioning of Nodes-B, any specific antenna pattern, and hard, soft, and softer handover mechanisms.

Integrated platform. We realized a flexible architecture integrating VISSIM with SHINE, enabling the realistic simulation of vehicular traffic together with real-time networks communications [8]. This integrated platform allows the simulation of the whole smart navigation scenario: the vehicular mobility and the OBUs's transmissions, the data processing at the control center, the data base update with dynamic data, and the retransmission of personalized dedicated information to the smart navigators. The architecture of the overall simulation platform is depicted in Fig. 2; each component interacts with the others through sockets and remote procedure calls (RCPs). In particular, the following main blocks can be identified: *OBU*, which simulates the on board device collecting and transmitting position and speed; *DRG*, which stands for dynamic rout guidance and

 ISBN: 978-1-61208-229-5

represents vehicles fleet equipped with a smart navigator device, able to receive updated traffic information from the control center; *Control Center*, which is responsible of the gathering of data transmitted by the OBUs to update the dynamic data base; *Communication Technology*, which simulates the cellular network from the application layer down to the physical layer; *Traffic Control Framework*, which is charge of controlling vehicles and their interactions and is based on VISSIM simulator; *Simulation Globals*, which manages the overall architecture from the traffic to the network simulation. More details are available in [8].

V. Service Classes and Figures of Merit

The road-network layout of the reference scenario consists of the medium sized Italian city of Bologna. In particular, we considered 13.636 road segments, corresponding to a length of about 600 Km. The digital-maps of the Italian road network have been provided by TeleAtlas, the world's provider of location and navigation solutions, and given as input to VISSIM. Results will be obtained considering two classes of service: vehicles equipped with OBUs performing packet transmissions over the cellular networks (hereafter intelligent transportation system (ITS) users), and pedestrians performing voice calls through cellular phones (hereafter voice users) as background traffic. Pedestrians can move everywhere in the scenario and are randomly generated in the scenario, with the same birth probability in each cellular cell (this means that a higher density of users is assumed where smaller cells are considered). Differently, vehicles' positions are managed by VISSIM.

A. Uplink

The portion of Bologna considered for extracting UMTS simulations results is shown in Fig. 3 and consists of a rectangular area of the city center sized 1.8 km (longitude) x 1.6 km (latitude) with 35 UMTS cells covered by 15 Nodes-B (1, 2 or 3 cells per Node-B are assumed). An approximated area of coverage is depicted for each cell with random colors. Black segments represent roads where vehicles movements are constrained. Hereafter, $\Lambda^{(v)}$ indicates the average offered voice load in Erlang per km^2, while $\Lambda^{(I)}$ indicates the average offered ITS load expressed in vehicles per Km^2. A single frequency planning is considered. In each cell, one RACH (out of the available ones) is exclusively used by the ITS service in the uplink. Propagation channels are realistically represented both for UMTS and IEEE 802.11p. Vehicles transmit 80 byte packets every 10 seconds. For further details the reader may refer to [4].

Figures of merit. To evaluate the quality of the ITS service, we aim at investigating the probability that each measurement stored in vehicles is correctly received by the control center, independently on the specific source.

A scheduled transmission fails in two cases: when the RACH ramping procedure is unsuccessful, meaning that the propagation conditions and the perceived interference level are so disadvantageous that the maximum transmission power is not sufficient, and when an error is checked at the receiver. In any case, the MAC layer may attempt a number of retransmissions before discarding the packet.

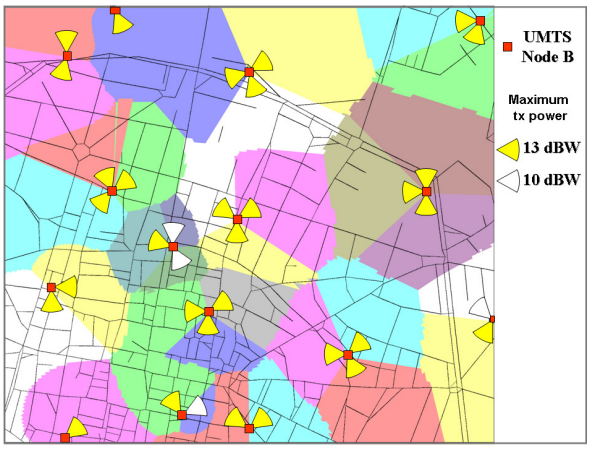


Figure 3. Map of the city center of Bologna and UMTS planning. Filled black squares correspond to Nodes-B locations. An approximated area of coverage is depicted for each cell with random colors.

Focusing on the ITS service, results will thus be expressed in terms of *packet discard rate* (R_D), that is the ratio between the number of discarded packets and the total number of packets generated by all on-board equipments.

B. Downlink

The portion of Bologna and the city planning of UMTS are those of the uplink described in Section V-A.

1) Unicast Mode: Following the assumptions made in Section III, vehicles receive updated traffic information through a 60000 bytes download (i.e., 1000 road segments $\times$ 60 bytes) every 5 minutes. Data are transmitted adopting the TCP protocol at the transport level, that assures data reception. A 64 kb/s bearer is considered, corresponding to a logical dedicated traffic channel (DTCH), a transport DCH, and a physical dedicated data channel (DPDCH).[1] The DPDCH is transmitted adopting a spreading factor (SF) equal to 16 in uplink (note, in fact, that a dedicated unicast communication is required also in the uplink direction for the TCP acknowledgment transmission) and 32 in downlink. A transmission time interval (TTI) of 10 ms is assumed.

Figure of merit. An ITS user is satisfied if the update is received with a delay lower than 15s (please consider that less than 10 seconds would be required if data were transmitted at 64 kb/s with no errors and no TCP redundancy).

2) Multicast Mode: Data are transmitted adopting the user datagram protocol (UDP) at transport level, which

[1]The low amount of bytes and the relaxed delay requirements do not justify the use of more consuming bearers.

 ISBN: 978-1-61208-229-5

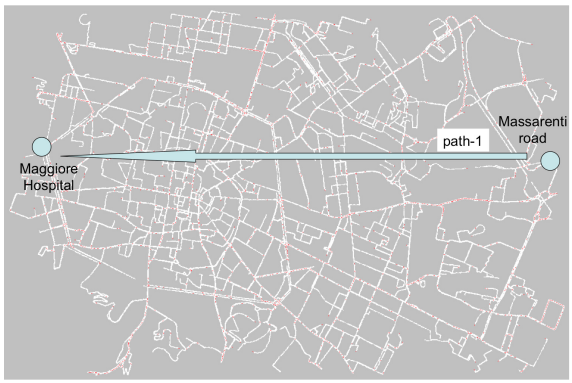


Figure 4. Origin and destination of path-1 in the considered scenario.

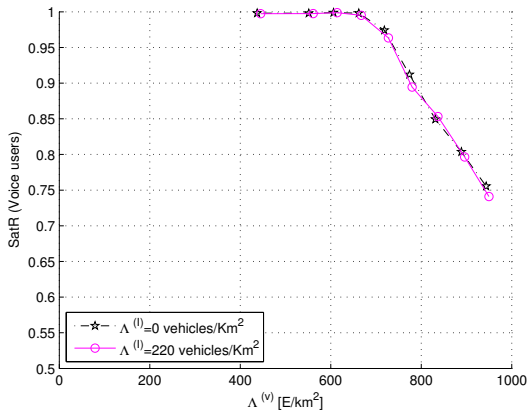


Figure 5. Voice traffic performance: voice *SatR* vs. the offered voice traffic $\Lambda^{(v)}$. Comparison between no ITS and ITS service on the network.

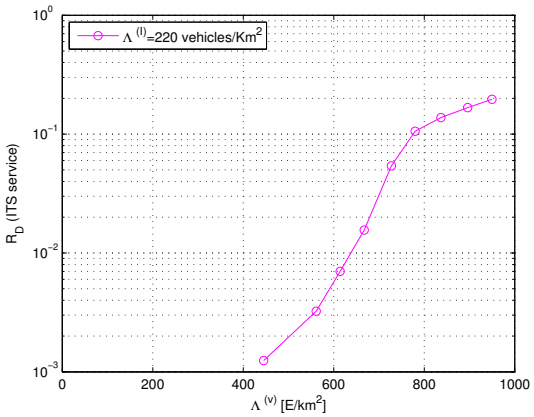


Figure 6. ITS traffic performance: ITS packets R_D varying the offered voice traffic $\Lambda^{(v)}$, with $\Lambda^{(I)} = 220$ vehicles/Km^2.

introduces limited redundancy but do not grant reliable communications; in this case, in fact, the absence of the uplink connection does not allow the transmission of acknowledgments. Two bearers at 64 and 128 kb/s are considered, each corresponding to an MTCH (MBMS transport channel) logical channel, a FACH (forward access channel) transport channel and a S-CCPCH (secondary common control physical channel) physical channel. The S-CCPCH is transferred (obviously, in downlink) adopting a SF equal to 64. A TTI of 40 ms is assumed.

Figure of merit. An ended ITS session is assumed in outage if less than the 95% of packets are correctly received.

C. Background Voice Traffic

To evaluate the UMTS performance in the considered scenario both in the uplink and in the downlink, the quality perceived by users belonging to other services than the ITS one is also of main interest. Without lacking of generality, here we focus on random walking users performing voice calls as interfered service both in the uplink and downlink.

Figures of merit. The evaluation of the quality of service perceived by users is based on the following definitions: per each frame lasting 10ms, a user (i.e., a voice call) is defined in outage if the BER after channel decoding of that frame is greater than 2% (uplink and downlink are evaluated independently to each other); an ended *voice call* is then considered *in outage* when either in downlink or in uplink, the outage intervals exceed a threshold of 5%. Hence, we have an outage voice call when one user is able to talk to the other party, but with poor audio quality.

A voice call may also incur in the following situations: it may be blocked by the call admission control algorithm due to insufficient resources, or it may drop due to an excessive reduction of the received signal power.

For this reason, results will be presented in the following in terms of *satisfaction rate (SatR)*, that is the ratio between the number of users which are not blocked, neither dropped, nor in outage, and the total number of call requests.

D. Short Range Communications

The use of V2V and V2R is envisioned in order to reduce both the cellular network load, which represents the main component of the service cost, and the delivery delays, that impact on the accuracy of vehicular traffic estimation (lower delays, in fact, mean a more frequent update of traffic conditions). In the simulations, a parametric percentage of vehicles is assumed equipped with an OBU that every τ seconds acquires several vehicle parameters, such as speed and position (which are referred in the following as *measured data*). Measured data are stored in the OBUs transmission queues and then transmitted according to the transmission strategies described in Section III-C. Both fluent traffic conditions and congested traffic conditions with car-queues arising in the proximity of some crossroads are considered. The former case is characterized by $\Lambda^{(I)} = 150$ vehicles/km^2 in average, whereas an average density of $\Lambda^{(I)} = 220$ vehicles/km^2 characterizes the latter case.

 ISBN: 978-1-61208-229-5

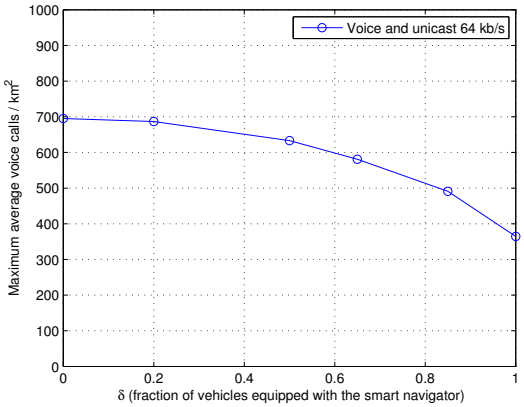


Figure 7. Voice capacity as a function of the fraction of vehicles receiving dedicated traffic information in unicast mode. $\Lambda^{(I)} = 220$ vehicles/Km^2.

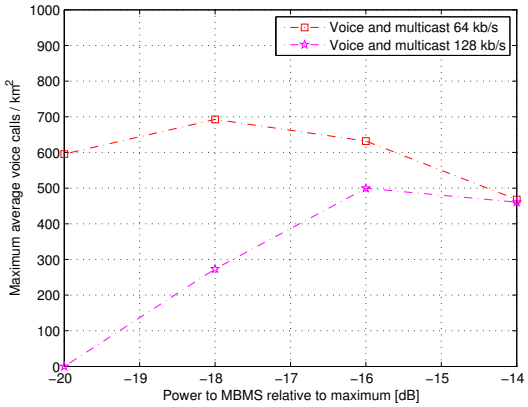


Figure 8. Voice capacity as a function of the Node-B power dedicated to MBMS multicast service. $\Lambda^{(I)} = 220$ vehicles/Km^2.

Curves will show the ratio of saved cellular resources S_R, i.e. the number of packets not delivered through cellular network exploiting the adoption of the considered strategy over the number of generated packets; considerations on delivery delay are not reported here for the seek of conciseness, and can be found with other details in [10].

E. Smart Navigation Service

To also evaluate the impact of a timely communication on traffic management, we evaluate the travel time to destination of a vehicle equipped with smart navigator. To this aim, the following are the simulation settings. In this case we considered non stationary traffic conditions, with a vehicles density that during each simulation vary in the range 1-10 vehicles/Km.

A parametric percentage of vehicles is assumed equipped with OBUs. No V2V and V2R communication is assumed in this case; every τ seconds, each OBU transmits the actual position and speed to the control center. Measured data are stored in the control center queue and averaged on a parametric T_{int} interval time.

Then, every T_{update} seconds, the control center retransmits the processed data back to those vehicles equipped with smart navigators. To avoid altered measurements in those roads where no vehicles or a too low percentage of them passed, we set up an average speed equal to that given by the static roadmap provided by TeleAtlas lowered by the 30%: this allows to not overestimate the speed. In addition, when the measured speed is lower than 15Km/h, we force the measurements exactly to 15Km/h: this avoid to overestimate the travel time in the involved road segment. As a case study, the origin-destination couple, denoted as *path-1* and represented in Fig. 4, has been considered in simulations.

VI. Numerical Results

A. Uplink

The *SatR* for voice users and the R_D for ITS service are plotted in Fig. 5 and Fig. 6, respectively, as a function of $\Lambda^{(v)}$. $\Lambda^{(I)} = 220$ vehicles/Km^2 is assumed, which corresponds to a heavy traffic condition with many traffic queues.

In Fig. 5, the *SatR* of voice users is depicted, and the case with no ITS service ($\Lambda^{(I)}$=0) is shown for comparison. Observing Fig. 5, the presence of the ITS service seems not to impact on voice users, since their satisfaction remains almost unchanged. In Fig. 6, however, the R_D as a function of $\Lambda^{(v)}$ is also plotted for the same value of $\Lambda^{(I)}$. As can be observed, the higher is the network load, the higher the R_D, and the QoS of the ITS service results deteriorated. If $\Lambda^{(v)}$=740 (corresponding to *SatR*=0.95) is taken as reference value, a packet loss higher than 5% can be observed, meaning that guaranteeing a *SatR*=0.95 to voice users, does not imply that the ITS users are also served. To improve the QoS of the ITS service, a lower number of voice calls must be accepted. For instance, if R_D lower than 10^{-2} is targeted, with respect to a maximum of $\Lambda^{(v)}$=740 in the absence of the ITS service, a reduction of about 100 (13.4%) average voice users per Km^2 must be considered, drastically reducing the voice users' capacity.

B. Downlink

In Fig. 7, the maximum voice capacity normalized in a one Km^2 area is plotted as a function of the number of equipped vehicles receiving updated traffic information via a dedicated unicast channel. In particular, the x-axis represents the ratio δ of vehicles that are equipped with the smart device. The y-axis represents the maximum amount of voice calls that allow the system to serve both traffic classes with a satisfaction rate (i.e., ratio of satisfied users over the number of users of that class) grater than 95%. When the number of equipped vehicles is zero, we obtain results referred to

 ISBN: 978-1-61208-229-5

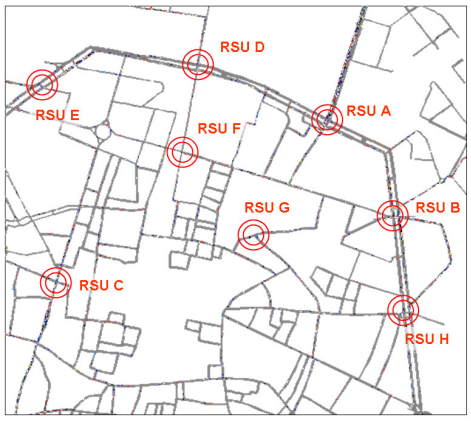


Figure 9. RSUs positions in the considered scenario.

the presence of voice only, considered as a benchmark (695 average voice calls). We can observe that, as the number of equipped vehicles receiving updated information increases, the maximum number of voice calls (i.e., the number of voice users) decreases, due to larger resources dedicated to the ITS service. However, we can note that, if the 50% of vehicles were equipped with smart navigators receiving updated information in unicast mode, the system could serve them also satisfying about 620 voice calls per Km^2. If all vehicles were equipped ($\delta = 1$), the capacity of the system in terms of servable voice calls would, instead, be halved.

Multicast results, shown in Fig. 8, are obtained varying the power used to transmit the S-CCPCH carrying the MTCH channel of the MBMS channel (which is used for the ITS service). More specifically, a constant fraction of the maximum available power at the base station is reserved for this use. In Fig. 8 results are presented assuming the fraction of Node-B power dedicated to MBMS in the x-axis and the maximum amount of voice calls (per km^2) that allow the system to serve both classes of traffic with at least 95% satisfaction rate in the y-axis. Multicast at 64 kb/s and 128 kb/s are compared. As can be observed, independently on the adopted bearer, the number of voice calls increases with the power dedicated to MBMS until a maximum, then it start decreasing: low power levels to the MBMS service, in fact, require low interference in order to guarantee a full coverage to the ITS service, while high power levels generate strong interference that limits the number of servable voice calls. We can thus note that a trade off between voice and ITS service can be obtained for both 64 kb/s and 128 kb/s, corresponding to -18 dB to MBMS with 690 average voice calls and -16 dB to MBMS with 500 average voice calls, respectively. These numbers also highlight that the adoption of a 128 kb/s bearer greatly reduces the number of voice calls with respect to 64 kb/s.

C. Short Range Communications

Numerical results are initially given considering a single RSU in the most crowded junction and then modifying the position and number of active RSUs. To increase the probability to reach the RSU before the time out for transmission over the cellular link expires, the maximum number of packets that can be queued is here set to a very high number (1000); with this assumption, in our results, the time out is always reached before the threshold on stored data.

In Fig. 9, the sites we considered for RSUs deployment in the reference scenario of Bologna are shown, corresponding to the mostly crowded junctions (note, in fact, that major junctions are suitable sites also owing to the likely presence of lighting, traffic lights, and therefore of power supply).

As first case we assume that only RSU A is available for communication in the entire scenario depicted in Fig. 9. RSU A is positioned, in particular, in the busiest crossroad of the whole scenario. The average ratio of saved cellular resources that can be achieved in this case taking advantage of both V2V and V2R communications is shown in Fig. 10 as a function of the sampling interval τ, varying the traffic conditions and the percentage of vehicles equipped with the OBU. Results show that even when a single RSU is properly positioned in a large area, a significant amount of data can be transmitted to it. Let us observe, in fact, that even with τ lower than 30 s and only the 10% of vehicles equipped with OBU, S_R is still significant (more than 10%). It could be verified that also delivery delays are reduced [10].

The impact of RSUs position on the benefit they can provide is quite relevant: in order to investigate this aspect we report, in Fig. 11, S_R considering different positions of a single RSU (RSU A, RSU B, and RSU C) and the deployment of all RSUs depicted in Fig. 9. For the sake of conciseness, only the case of heavy traffic and 100% equipped vehicles is considered. Curves are shown as a function of the sampling interval τ. We can observe that RSU B and RSU C, being located in less busy junctions, are less effective than RSU A, owing to a reduced amount of vehicles passing in their proximity. Comparing the benefit provided by RSU A to the one achievable with all RSUs simultaneously active we can infer that the advantage obtained deploying all RSUs is not very high with respect to the one achievable deploying only RSU A and would not justify the financial investments required. The great effectiveness of RSU A suggests that a relevant role is played by V2V communications. It is likely, in fact, that also vehicles not passing in the proximity of RSU A are successful in the attempt to transmit their information without cellular transmissions by means of multi-hop V2V communications and the strategic position of RSU A. Here we can observe that the effectiveness of the roadside infrastructure is significantly enhanced by the joint adoption of V2V communications. If V2V interface was not available, then the number of deployed RSUs would

 ISBN: 978-1-61208-229-5

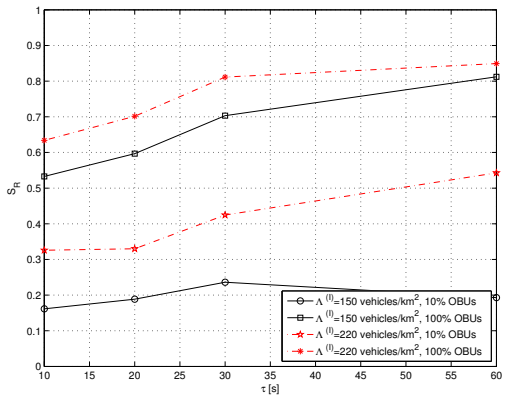


Figure 10. S_R as a function of τ, with RSU A only.

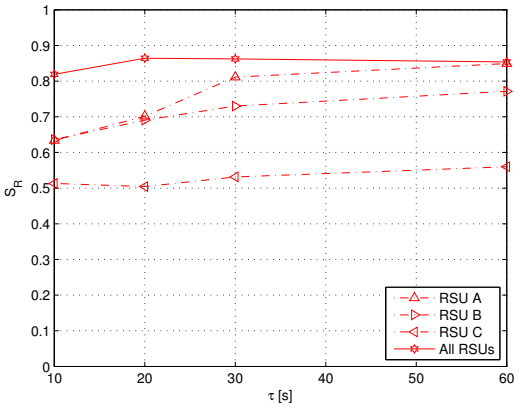


Figure 11. Impact of RSUs position and number. S_R as a function of τ with $\Lambda^{((I))} = 220$ vehicles/Km^2 and 100% vehicles equipped with OBUs.

be decisive to reduce the cellular load.

D. Smart Navigation Service

Results are presented for $T_{\text{int}} = 10$ s, $T_{\text{update}} = 20$ s, and τ equal to 10, 30, or 60 s. Figures show the travel time, from origin to destination, of a controlled vehicle equipped with the smart navigator for different percentage of equipped vehicles and different scenarios. In each figure, results are compared with the following three cases adopted as benchmarks: i) *Free running*, referred to the case of a single vehicle moving alone on the entire scenario; ii) *Best case with smart navigation*, referred to a vehicle equipped with a smart navigator continuously updated with the best route, iii) *No smart navigation*, referred to the the same route as in Free Running in the presence of traffic (a navigator may be present, but without knowledge of real time traffic).

In Fig. 12, the travel time to destination is shown for path-1. These results follow an extensive simulation campaign, where also other paths were considered; similar results have been however obtained in all cases, and other plots are here omitted for length limitation. Figures 12(a), 12(b), and 12(c) refer to three different (uplink) transmission times τ: 10 s, 30 s, and 60 s, respectively. For each percentage of OBU equipped vehicles, six results are presented, corresponding to six different simulations providing time and space randomness (i.e., different vehicles are equipped with OBU, and the sampling process starts at different instants). By observing Figs. 12(a), 12(b), and 12(c), it can be noted that the time to destination increases with τ, showing a not negligible impact of a timeliness update of road segments status. Focusing on Fig. 12(a), the impact of the percentage of OBU equipped vehicles can be appreciated: with a so prompt update in the uplink ($\tau = 10$ s), the 10% of vehicles equipped with connected OBUs is sufficient to have a time to destination very near to the best case (i.e., about 600 s). When the transmission time τ from vehicles to the control center is higher (see Figs. 12(b), and 12(c) referring to $\tau = 30$ s and $\tau = 60$ s, respectively) the 10% of vehicles is no more sufficient to obtain optimal results, that can be instead achieved only when all vehicles are equipped with OBUs.

VII. Conclusions

In this work, we summarized the results of the Italian project PEGASUS with focus on wireless technologies for smart mobility. Specifically, we considered UMTS as the enabling technology for the real time acquisition and transmission of traffic information. We discussed the feasibility of the service and we evaluated the impact of such a communication on other services already provided by UMTS. Our studies highlighted that the service appears feasible and that the number of equipped vehicles does not seem a critical issue; we also pointed out, however, that a not negligible loss in capacity for the other services must be accounted for in order to guarantee a satisfactory quality of service. The benefits arising from the adoption of V2V and V2R communications have also been explored. We finally showed that the knowledge in real time of the traffic conditions allows an efficient smart navigation and a not negligible saving of travel time.

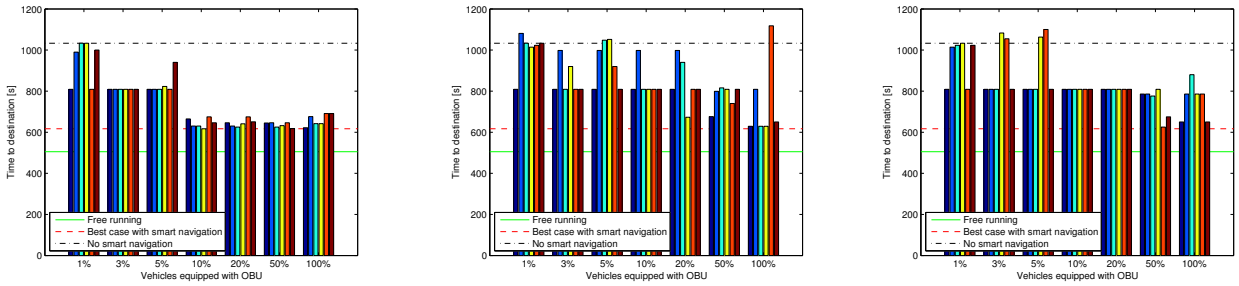
References

[1] P. Papadimitratos, A. La Fortelle, K. Evenssen, R. Brignolo, and S. Cosenza, "Vehicular communication systems: Enabling technologies, applications, and future outlook on intelligent transportation," *Communications Magazine, IEEE*, vol. 47, no. 11, pp. 84 –95, 2009.

[2] *IEEE Trial-Use Standard for Wireless Access in Vehicular Environments (WAVE) - Multi-Channel Operation*, Std., 2006.

[3] *Standard for Information Technology- Telecommunications and Information Exchange between Systems- Local and Metropolitan Area Networks-Specific Requirements Part 11 - Amendment 6: Wireless Access in Vehicular Environment*, IEEE Std. 802.11p, 2010.

 ISBN: 978-1-61208-229-5



(a) $\tau = 10$s. (b) $\tau = 30$s. (c) $\tau = 60$s.

Figure 12. Travel time to destination for several percentage of vehicles equipped with OBU.

[4] A. Bazzi, B. Masini, and O. Andrisano, "On the frequent acquisition of small data through RACH in UMTS for ITS applications," *Vehic. Tech., IEEE Tran. on*, vol. 60, no. 7, pp. 2914 – 2926, Sep. 2011.

[5] PEGASUS project. Accessed on Oct. 2012. [Online]. Available: http://pegasus.octotelematics.com

[6] A. Conti, A. Bazzi, B. Masini, and O. Andrisano, *Vehicular Networks: Techniques, Standards, and Applications*. Auerbach Pub., CRC Press, Taylor & Francis Group, 2009, ch. Heterogeneous Wireless Communications for Vehicular Networks, pp. 63–107, ISBN: 9781420085716.

[7] A. Bazzi, B. Masini, G. Pasolini, and P. Torreggiani, "Telecommunication systems enabling real time navigation," in *IEEE ITSC*, Sep. 2010, pp. 1057 – 1064.

[8] A. Toppan, A. Bazzi, P. Toppan, B. Masini, and O. Andrisano, "Architecture of a simulation platform for the smart navigation service investigation," in *IEEE WiMob*, Oct. 2010, pp. 548 – 554.

[9] A. Bazzi and B. Masini, "Real time traffic updates via UMTS: Unicast versus multicast transmissions," in *IEEE VTC Fall,*, Sep. 2011, pp. 1 – 6.

[10] A. Bazzi, B. Masini, and G. Pasolini, "V2V and V2R for cellular resources saving in vehicular applications," in *IEEE WCNC*, Apr. 2012, pp. 1–5.

[11] P. Belanovic, D. Valerio, A. Paier, T. Zemen, F. Ricciato, and C. Mecklenbrauker, "On wireless links for vehicle-to-infrastructure communications," *Vehic. Tech., IEEE Trans. on*, vol. 59, no. 1, pp. 269 – 282, Jan. 2010.

[12] D. Valerio, A. D'Alconzo, F. Ricciato, and W. Wiedermann, "Exploiting cellular networks for road traffic estimation: A survey and a research roadmap," in *IEEE VTC Spring 2009*, april 2009, pp. 1 –5.

[13] The aktiv CoCar project. Accessed on Oct. 2012. [Online]. Available: http://www.aktiv-online.org/english/aktiv-cocar.html

[14] S. Cho, K. Geon, Y. Jeong, C.-H. Ahn, S. I. Lee, and H. Lee, "Real time traffic information service using terrestrial digital multimedia broadcasting system," *Broadcasting, IEEE Transactions on*, vol. 52, no. 4, pp. 550 –556, Dec. 2006.

[15] U. Dietz, "CoCar Feasibility Study: Technology, Business and Dissemination," CoCar Consortium, Report, May 2009.

[16] C. Sommer, A. Schmidt, R. German, W. Koch, and F. Dressler, "Simulative Evaluation of a UMTS-based Car-to-Infrastructure Traffic Information System," in *IEEE GLOBECOM*. New Orleans, LA: IEEE, December 2008.

[17] C. Sommer, A. Schmidt, Y. Chen, R. German, W. Koch, and F. Dressler, "On the feasibility of umts-based traffic information systems," *Ad Hoc Networks*, vol. 8, no. 5, pp. 506 – 517, 2010.

[18] I. Lequerica, P. Ruiz, and V. Cabrera, "Improvement of vehicular communications by using 3G capabilities to disseminate control information," *Network, IEEE*, vol. 24, no. 1, pp. 32 – 38, Jan.-Feb. 2010.

[19] B. Bai, W. Chen, K. Letaief, and Z. Cao, "Low complexity outage optimal distributed channel allocation for vehicle-to-vehicle communications," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 1, pp. 161 –172, 2011.

[20] J. Karedal, N. Czink, A. Paier, F. Tufvesson, and A. F. Molisch, "Path loss modeling for vehicle-to-vehicle communications," *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 1, pp. 323 –328, 2011.

[21] C. Campolo and A. Molinaro, "Vehicle-to-roadside multihop data delivery in 802.11p/WAVE vehicular ad hoc networks," in *IEEE GLOBECOM*, 2010, pp. 1 –5.

[22] P. Benko, G. Malicsko, and A. Veres, "A large-scale, passive analysis of end-to-end TCP performance over GPRS," in *IEEE INFOCOM*, vol. 3, Mar. 2004, pp. 1882–1892.

[23] I. Vukovic and T. Brown, "Performance analysis of the random access channel (RACH) in WCDMA," in *IEEE VTC Spring.*, vol. 1, May 2001, pp. 532–536.

[24] *Traffic and Travel Information (TTI) via Transport Protocol Expert Group (TPEG) data-streams: Parts 1 to 6*, ETSI Std. ISO 18 234.

[25] Vissim. Accessed on Oct. 2012. [Online]. Available: http://www.english.ptv.de/cgi-bin/traffic/traf_vissim.pl

[26] A. Bazzi, G. Pasolini, and C. Gambetti, "SHINE: Simulation platform for heterogeneous interworking networks," in *IEEE ICC*, vol. 12, 2006, pp. 5534 –5539.

 ISBN: 978-1-61208-229-5

VVID: A Delay Tolerant Data Dissemination Architecture for VANETs Using V2V and V2I Communication

Koosha Paridel, Yolande Berbers
Department of Computer Science
KU Leuven
Leuven, Belgium
Email: {koosha.paridel, yolande.berbers}@cs.kuleuven.be

Josip Balen, Goran Martinovic
Faculty of Electrical Engineering
J. J. Strossmayer University of Osijek
Osijek, Croatia
Email: {josip.balen, goran.martinovic}@etfos.hr

***Abstract*—Vehicular Ad-Hoc Networks (VANETs) have a highly dynamic network topology due to the constant and rapid movement of vehicles. Therefore, communication in VANETs is mostly affected with disruptions and delays as a result of network disconnections and partitions. Moreover, increasing the geographical coverage area for data dissemination while keeping the data-delivery delay low is a challenge. This paper first surveys the routing techniques in Delay Tolerant Networks (DTNs), and also the usage of DTN communication in VANETs. Then, the paper proposes a combined architecture of Vehicle-to-Vehicle (V2V) communication, DTN communication, and Vehicle-to-Infrastructure (V2I) communication as a large-scale data dissemination system for vehicular networks. We argue that the combined architecture enables data dissemination in a larger geographical area compared to a V2V communication model, and efficiently deals with disconnectivity and network partitions that mostly occurs in sparse networks.**

Keywords-VANET; DTN; V2V Communication; V2I Communication.

I. INTRODUCTION

Wireless communication in Vehicular Ad-Hoc Networks (VANETs) enables information exchange between vehicles (V2V communication) and between vehicles and roadside infrastructure (V2I communication). Since vehicles are moving, network topology is constantly changing. Therefore, VANETs are highly dynamic and have most characteristics of Mobile Ad-Hoc Networks (MANETs) [1]. However, VANETs behave in different ways than conventional MANETs, since vehicle movements are constrained by roads and mobility patterns can be predicted to some extent. Furthermore, vehicles have different characteristics than conventional nodes in MANETs. There are numerous applications in the domain of VANETs and they can be used for different purposes like improving safety (accident avoidance, incident notification), improving driving (congestion monitoring, parking space allocation) and commercial services (business, entertainment).

Our current work, described in [2], optimizes communication in VANETs with context-based grouping mechanism based on common spatio-temporal characteristics (location, direction, speed and time) and shared interests. Each group is represented by vehicles with special responsibilities that specify which context information can be distributed inside the group and between the groups. The large-scale simulated experiments show that our context-based grouping mechanism significantly reduces the overall network traffic usage, irrelevant and redundant information flow and the processing overhead.

Our experiment on very large-scale realistic traffic data shows that despite the benefits of group-based communication, there is still a considerable number of vehicles that are far away from the crowded areas and are not in contact with other vehicles. These vehicles form single-vehicle groups and while they are disconnected from the rest of the traffic, they cannot participate in the propagation of information. These vehicles can come in contact with other vehicles later on, but they lose all the information propagated during their disconnection.

Taking advantage of temporary connections to propagate information is the focus of delay-tolerant communication. A delay-tolerant network is composed of nodes with bidirectional links together. These links may disconnect and connect as a result of mobility or failure. When two nodes are connected, they have the opportunity to exchange data. In a delay-tolerant communication system, nodes, instead of only receiving and forwarding the information, also save the information in a buffer to exchange it later on, when they come in contact with other nodes. This model is usually referred to as the *store-carry-forward* model. This characteristic of the delay-tolerant communication matches with what our group-based communication is lacking. Adding delay-tolerant communication capability helps our current system to increase its information propagation coverage to less crowded areas where there is very limited connection.

The contribution of the paper is a hybrid architecture called VVID that combines V2V, V2I, and Delay Tolerant Network (DTN) communication models in order to address the current issues in vehicular communication such as geographical scalability, information coverage in sparse networks, and smart and timely dissemination of information.

The rest of the paper is organized as follows: Section II describes the routing techniques in DTNs. In Section III, we survey recent work in the area of DTN communication in VANETs. In Section IV, we propose and describe a hybrid architecture for the efficient data dissemination in VANETs that combines V2V, V2I and DTN communication models. Section V concludes the paper.

 ISBN: 978-1-61208-229-5

II. ROUTING TECHNIQUES IN DTNS

As mentioned in [3], DTNs are often described with various phrases, such as eventual, partially, intermittently or transient connected networks, opportunistic networking, and space-time routing. All aforementioned terminologies are used to describe a network where end-to-end connectivity is not assumed and communication is affected with disruptions and delays as a result of network disconnections and partitions. The best example of DTNs are MANETs since nodes in such networks are usually moving and network structure is constantly changing. To overcome all mentioned shortcomings, efficient routing protocols are required. As described in [3], based on network time-evolving topology they can be categorized as routing protocols for deterministic or stochastic time-evolving networks. If all future topologies of the network are completely known or predictable then deterministic routing can be applied. There are three different approaches: (i) Space time routing; (ii) Tree approach; and (iii) Modified shortest path approach. They are all based on modeling the dynamics of the network as a space-time graph or tree and then selecting the final path depending on requirements (shortest time or minimum number of hops). In dynamic networks, where network behavior is random, future network topology is unknown and cannot be predictable. Therefore, to deliver packets from source to destination routing protocols for stochastic time-evolving networks should be applied. There are five different approaches:

- Epidemic routing-based approach [4], [5], [6], [7]: The classic example is to flood the message through the network. However, this approach is very costly and can cause network congestion. Another example is to deliver message only when source and destination are within communication range. Although this approach has minimal overhead, the delay is often very long. The best results are obtained with approaches that are trade-off between these two extreme examples. Another example is spray routing. The idea is to first unicast sprayed packet to a node close to destination and afterwards multicast traffic within the vicinity of the last-known location of the destination. Furthermore, by adding relay nodes between source and destination routing performances can be significantly improved.
- History or predication-based approach [8], [9]: In this approach the link forwarding probability is estimated based on the one-hop or end-to-end information. Nodes can interrogate each other to learn more about network topology and nodal capacity to make intelligent routing decisions. One-hop information is usually obtained by exchanging the information between two nodes when they meet. The selection of the next hop can be based on the various metrics, such as: spatial location, bandwidth, relative velocity/mobility between two nodes, vicinity of the candidate, capability of the candidate and data transmission time.
- Model based approach [10]: It is based on modeling motion patterns of mobile nodes for a better selection of relaying nodes and a determination of receiver's location without flooding the network. For example, in VANETs, the two mostly used vehicle traffic models are: highways and city traffic models.
- Node movement control-based approach [11] [12]: Controlling node mobility can improve overall system performance. There are different ways to control node mobility such as modifying nodes trajectories, using virtual mobile nodes that travel through the network and collect and deliver messages, controlling the mobility of autonomous agents, using Message Ferries (single or multiple) to provide communication services for nodes in the network, using snake and runners protocols and using DataMules.
- Coding based approach [13], [14]: Erasure coding and network coding techniques are used in this approach. In erasure coding technique an original message is encoded into a large number of coding blocks but can be decoded if smaller number of blocks is received. With this technique worst case delay can be significantly improved. In network coding, intermediate nodes, instead of simply forwarding the packets they receive, can combine some received packets and send them out as a new packet. By using this technique, packet delivery ratio could be much higher.

III. DTN COMMUNICATION IN VANET

There are several works that apply DTN routing techniques in VANETs. Yang and Chuah [15] presented Ferry Based Interdomain Multicast Routing Scheme (FBIMR) in DTNs where ferries are used to deliver multicast messages across groups that are partitioned. They are investigating how different buffer sizes, numbers of ferries and ferry speeds impact on the delivery performance in VANETs. They concluded that by increasing the buffer size (from 1000 to 2500 packets) the delivery ratio is improved but the average delay is increased and data efficiency is decreased. Furthermore, when the packet rate is higher than 1 pkt/s delivery ratio is significantly increased and delay is decreased. With increasing the ferry speed from 15 m/s to 30 m/s delivery ratio is much higher and delay is lower.

Zhao et al. propose an infrastructure-to-vehicles data dissemination system with a buffering mechanism to increase the data dissemination coverage [16]. In their system, there are several data centers installed along a road. These data centers periodically broadcast data to the vehicles along the road. This procedure is called Data Pouring (DP). The vehicles that receive a broadcast, buffer the data and re-broadcast it in the intersections. Therefore, the system covers not only the road covered with data centers, but also the intersecting roads. The authors call this mechanism as *DP with Intersection Buffering (DP-IB)*.

VADD (Vehicle-Assisted Data Delivery in

 ISBN: 978-1-61208-229-5

VANET) [17] adopts the idea of store-carry-forward communication model to deal with the intermittent connectivity in VANETs. Moreover, VADD tries to predict the mobility of the vehicles in order to forward packets to the best route with the lowest data delivery delay. In VADD, vehicles store and carry the information, and forward them in the intersections. Based on the technique used for road selection at the intersections, the authors propose three different VADD protocols, and evaluate them in terms of packet-delivery ratio, data packet delay and traffic overhead.

DV-CAST [18] is a vehicular broadcast protocol that aims to address both the broadcast storm problem [19] and the disconnected network problem, in order to build a system that works efficiently in dense and sparse network areas. DV-CAST consists of three main components, namely neighbor detection, broadcast suppression, and store-carry-forward mechanisms. Neighbor detection mechanism provides the local connectivity information, and on the basis of this information, broadcast suppression mechanism or store-carry-forward mechanism is used to deal with the dense or the sparse network situations respectively. DV-CAST is evaluated only in highway scenarios and the authors are planning to extend their solution to urban environments as well.

SRD [20] is a dissemination protocol for VANETs, which employs different strategies for dense and sparse networks. SRD uses broadcast suppression techniques in dense networks, in order to avoid the broadcast storm problem. In sparse networks, SRD uses store-carry-forward communication model, in order to take advantage of the mobility of the nodes, and disseminate information in parts of the network that are geographically separated. The authors show through simulations that SRD outperforms DV-CAST [18] in terms of delivery ratio and network traffic overhead.

GeoDTN+Nav [21] is another hybrid routing protocol, which combines geographic routing and DTN forwarding. In dense or connected VANETs it uses geographic routing and routes packets in two modes: greedy and perimeter mode. It is able to estimate network partition and then improves packet delivery by switching to the DTN mode. Furthermore, authors propose Virtual Navigation Interface (VNI) that provides generalized route information in order to choose forwarders in partitioned networks. Performance evaluation shows that GeoDTN+Nav outperforms conventional geographic protocols in packet delivery ratio. However, the tradeoff is an increased delivery delay.

Bitaghsir and Hendessi [22] proposed an intelligent routing protocol for DTNs, which is similar to the GeoDTN+Nav protocol presented in [21]. For the dense or connected VANETs they both use the same geographic routing. However, for the delay tolerant forwarding the intelligent routing protocol considers other useful parameters in order to choose the best node for storing and carrying the packets. Furthermore, the genetic algorithm is used to train the DNT node evaluation system and to determine how important each parameter is in the simulation environment. Performance evaluation results show that each new generation of parameters obtained by the genetic algorithm decreases the average delivery delays and increases average delivery ratio.

GeoSpray [23] is a hybrid geographic routing protocol that takes advantages of multiple-copy and single-copy routing schemes. First it performs ”control spraying” by distributing a limited number of bundle copies to the network nodes that go closer (and/or arrive sooner) to the bundle destination. Afterwards, it switches to a forwarding scheme where it combines several control data sources to perform routing decisions. In the end, in order to improve resource utilization, it clears delivered bundles across the network nodes. Performance measurement results show that it improves the delivery probability and reduces delivery delay, comparing with the other multiple-copy and single-copy routing schemes.

While the mentioned systems use DTN communication to cover the sparse areas of network, none of these systems provide a solution for communication in geographically large vehicular networks. The simulations are done in areas with dimensions of a few kilometers, and the number of vehicles is limited to a few hundred. In order to cover geographical areas with dimensions of hundreds of kilometers, one cannot only rely on multihop V2V communication, as it increases the propagation delay, decreases the delivery ratio, and imposes heavy overhead, since the vehicles must buffer many messages, keep them for a log period and carry them for large distances.

IV. Problem Statement and Proposed Solution

In our previous work [2], we took advantage of the network of vehicles, and used this ad hoc network in order to disseminate important information among vehicles. We evaluated our mechanism using simulations. We measured parameters such as network traffic, message delivery effort and relevancy of the delivered information. However, we encountered several limitations while evaluating our mechanism using simulations. For example, it was not possible to simulate large datasets because it requires considerable amount of time, resources and computational power. Furthermore, due to the lack of a global view over the ad hoc networks, it is difficult to analyze the structure of the groups and their connectivity in the network using simulation.

Therefore, we developed a tool using Prolog for analysis of vehicular ad hoc networks. This tool receives the traces of movements of the vehicles for a certain time period as an input. On the basis of the logic of a routing protocol, our tool measures the desired network parameters. By using our tool, we can now measure the number of groups at each snapshot of the time. We also measure parameters such as the average number of vehicles in each group, the number of connections between the groups, the shortest paths between groups and the graph diameter of the network. An important contribution of this work is the ability to analyze a potentially large network with a massive number of vehicles such as a VANET. In our

 ISBN: 978-1-61208-229-5

experiment, we use a real-life vehicular dataset [24] with 260,000 vehicles recorded over a period of 24 hours.

An important observation during the analysis of a large-scale vehicular network was the existence of many single vehicles at each snapshot. These single vehicles are the ones that are moving in sparse parts of the network, and do not have any connection to the rest of the vehicles in the network. These vehicles can later on move into the denser areas and come in contact with other vehicles. However, these vehicles lose the opportunity of obtaining some important information while they are not connected to the rest of network.

The intermittently connected nature of vehicular networks requires a communication mechanism that can tolerate disconnectivities and delays more than the conventional IP delays. This is the main motivation behind delay tolerant communication. We believe that adding delay tolerant communication capability to our system can improve the coverage of information dissemination, and prevents losing important information caused by temporary disconnectivities.

We define a set of requirement for a data dissemination system in vehicular networks:

- **Coverage of Information**: We want to disseminate information to as many vehicles as possible, in a large geographical area of hundreds of kilometers, including areas with low density of vehicles.
- **Timeliness**: Fast dissemination of urgent information in a matter of seconds. Certain types of information such as reports of an accident or a hazard must be disseminated quickly in the few-kilometer vicinity of the incident.
- **Smart Dissemination**: Disseminate information based on the interest of vehicles. Prevent the dissemination of irrelevant information.
- **Minimal infrastructure**: The system must use as less infrastructure as possible. Infrastructure costs money, and is a point of failure, in contrast with infrastructure-less systems that are mostly cheap and more fault tolerant.

We propose an architecture for data dissemination in vehicular networks to address the above requirements. The architecture uses V2V communication, DTN communication, and V2I communication, combined together to build a system that can cover a big geographical area, handle disconnections and communication disruptions, and disseminate urgent information with a small delay. Figure 1 shows a simplified overview of our proposed architecture.

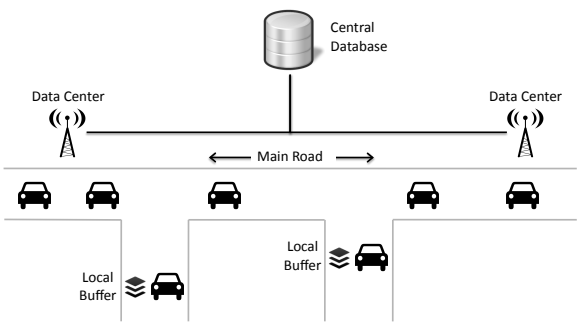


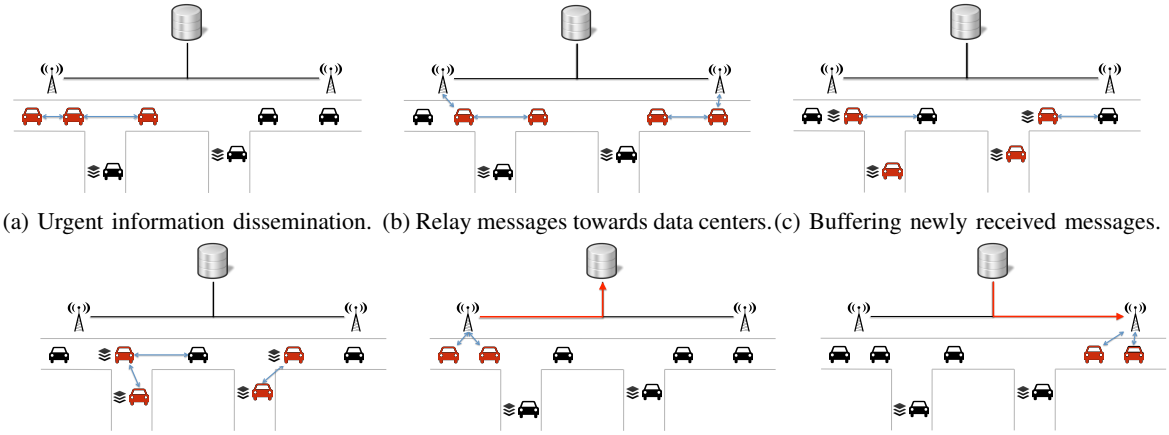
Figure 1. Main elements of the proposed architecture for data dissemination in vehicular networks.

We summarize the tasks of the three communication models used in our architecture as follows:

- **V2V Group-based communication**
  - **Share urgent information**: In case of emergency, such as reporting an accident, the vehicles use V2V communication to propagate the information in their vicinity, up to a certain hop limit (e.g., 10 hops). Therefore, the urgent messages are propagated with the lowest delay in a limited geographical area (see Figure 2a).
  - **Relay other information towards data centers**: Vehicles use V2V communication to relay non-urgent information towards data centers, so that the data centers can store them in the central database (see Figure 2b).
- **DTN communication**
  - **Buffer received data**: Vehicles store the received information in their buffers (see Figure 2c), and keep the buffer up-to-date by removing the old information.
  - **Exchange in case of contact with other vehicles**: When vehicles come in contact with each other, they compare each others buffer, and exchange the information that each vehicle is missing. Thus, the vehicles that are disconnected from the network of vehicle have the chance to receive the disseminated information with a delay (see Figure 2d).
- **V2I communication**
  - **Store the data collected by vehicles**: Data centers receive the propagated information in the network of vehicles and store them in the central database (see Figure 2e).
  - **Rebroadcast in other data centers based on the area of relevance**: Depending on the context, the stored information in the database can be rebroadcasted by the other centers, in other geographical areas (see Figure 2f).

In the VVID architecture, there are several data centers installed along the main highways that are all connected to a central database. These data centers do not have to cover the whole highway. They are only placed at specific parts of the highways with major traffic flow. When the vehicles move into the covered area of a data center, they can communicate with the data center and share information. The vehicles can also communicate together and share information without using the data center, according to our multi-hop group-based data dissemination protocol [2]. Moreover, the vehicles can buffer data to form a history of the recent information they have received. In sparse parts of the network, there are not enough vehicles to form a connected network of vehicles, and therefore, multi-hop communication is not effective without a store-carry-

 ISBN: 978-1-61208-229-5



(a) Urgent information dissemination. (b) Relay messages towards data centers. (c) Buffering newly received messages.

(d) History exchange upon contact with other vehicles. (e) Sending messages from vehicles to data centers. (f) Broadcasting the collected data in other data centers.

Figure 2. The combination of V2V, V2I and DTN communication model used in VVID architecture.

forward technique. In the sparse areas, vehicles exchange their received information history with other vehicles upon contact, in order to disseminate information to the vehicles that are disconnected from the rest of the network.

V. Conclusion and Future Work

Efficient data dissemination in vehicular networks is a challenge. A desired dissemination system must propagate urgent information in a timely manner, cover a large geographical area, cover the sparse part of the network, and prevent the propagation of irrelevant information. We propose a hybrid architecture for efficient data dissemination in vehicular networks called VVID that combines V2V, V2I and DTN communication models, in which pure V2V communication is used to propagate urgent information, a combination of V2V and V2I is used to propagate information in a large geographical area, and DTN communication to cover the sparse parts of network.

As future work, we will evaluate the proposed system in terms of network traffic, delivery delay, and relevance of information using simulation with realistic vehicular traces.

References

[1] Y. Toor, P. Muhlethaler, and A. Laouiti, "Vehicle ad hoc networks: applications and related technical issues," *Communications Surveys Tutorials, IEEE*, vol. 10, no. 3, pp. 74 –88, quarter 2008.

[2] K. Paridel, T. Mantadelis, A.-U.-H. Yasar, D. Preuveneers, G. Janssens, Y. Vanrompay, and Y. Berbers, "Analyzing the efficiency of context-based grouping on collaboration in vanets with large-scale simulation," *Journal of Ambient Intelligence and Humanized Computing*, vol. 3, pp. 1–16, May 2012.

[3] Z. Zhang, "Routing in intermittently connected mobile ad hoc networks and delay tolerant networks: overview and challenges," *IEEE Communications Surveys Tutorials*, vol. 8, no. 1, pp. 24–37, 2006.

[4] A. Vahdat and D. Becker, "Epidemic routing for partially-connected ad hoc networks," Duke University, Tech. Rep. CS-200006, Apr. 2000.

[5] M. Grossglauser and D. N. C. Tse, "Mobility increases the capacity of ad hoc wireless networks," *IEEE/ACM Trans. Netw.*, vol. 10, no. 4, pp. 477–486, Aug. 2002. [Online]. Available: http://dx.doi.org/10.1109/TNET.2002.801403

[6] D. Nain, N. Petigara, and H. Balakrishnan, "Integrated routing and storage for messaging applications in mobile ad hoc networks," *Mob. Netw. Appl.*, vol. 9, no. 6, pp. 595–604, Dec. 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1035715.1035720

[7] F. Tchakountio and R. Ramanathan, "Tracking highly mobile endpoints," in *Proceedings of the 4th ACM international workshop on Wireless mobile multimedia*, ser. WOWMOM '01. New York, NY, USA: ACM, 2001, pp. 83–94. [Online]. Available: http://doi.acm.org/10.1145/605991.606003

[8] M. Musolesi, S. Hailes, and C. Mascolo, "Adaptive routing for intermittently connected mobile ad hoc networks," in *Proceedings of the Sixth IEEE International Symposium on World of Wireless Mobile and Multimedia Networks*, ser. WOWMOM '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 183–189. [Online]. Available: http://dx.doi.org/10.1109/WOWMOM.2005.17

[9] C.-C. Shen, G. Borkar, S. Rajagopalan, and C. Jaikaeo, "Interrogation-based relay routing for ad hoc satellite networks," in *Global Telecommunications Conference, 2002. GLOBECOM '02. IEEE*, vol. 3, nov. 2002, pp. 2920 – 2924 vol.3.

[10] Z. D. Chen, H. Kung, and D. Vlah, "Ad hoc relay wireless networks over moving vehicles on highways," in *Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking & computing*, ser. MobiHoc '01. New York, NY, USA: ACM, 2001, pp. 247–250. [Online]. Available: http://doi.acm.org/10.1145/501449.501451

 ISBN: 978-1-61208-229-5

[11] S. Dolev, S. Gilbert, N. A. Lynch, E. Schiller, A. A. Shvartsman, and J. L. Welch, "Virtual mobile nodes for mobile ad hoc networks," in *in DISC04*, 2004, pp. 230–244.

[12] R. C. Shah, S. Roy, S. Jain, and W. Brunette, "Data mules: Modeling a three-tier architecture for sparse sensor networks," in *In IEEE SNPA Workshop*, 2003, pp. 30–41.

[13] S. Jain, M. Demmer, R. Patra, and K. Fall, "Using redundancy to cope with failures in a delay tolerant network," in *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, ser. SIGCOMM '05. New York, NY, USA: ACM, 2005, pp. 109–120. [Online]. Available: http://doi.acm.org/10.1145/1080091.1080106

[14] Y. Wang, S. Jain, M. Martonosi, and K. Fall, "Erasure-coding based routing for opportunistic networks," in *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, ser. WDTN '05. New York, NY, USA: ACM, 2005, pp. 229–236. [Online]. Available: http://doi.acm.org/10.1145/1080139.1080140

[15] P. Yang and M. Chuah, "Efficient interdomain multicast delivery in disruption tolerant networks," in *Mobile Ad-hoc and Sensor Networks, 2008. MSN 2008. The 4th International Conference on*. IEEE, 2008, pp. 81–88.

[16] J. Zhao, Y. Zhang, and G. Cao, "Data pouring and buffering on the road: A new data dissemination paradigm for vehicular ad hoc networks," *Vehicular Technology, IEEE Transactions on*, vol. 56, no. 6, pp. 3266–3277, 2007.

[17] J. Zhao and G. Cao, "Vadd: Vehicle-assisted data delivery in vehicular ad hoc networks," *Vehicular Technology, IEEE Transactions on*, vol. 57, no. 3, pp. 1910–1922, 2008.

[18] O. Tonguz, N. Wisitpongphan, and F. Bai, "Dv-cast: A distributed vehicular broadcast protocol for vehicular ad hoc networks," *Wireless Communications, IEEE*, vol. 17, no. 2, pp. 47–57, 2010.

[19] Y. Tseng, S. Ni, Y. Chen, and J. Sheu, "The broadcast storm problem in a mobile ad hoc network," *Wireless networks*, vol. 8, no. 2, pp. 153–167, 2002.

[20] R. S. Schwartz, R. R. R. Barbosa, N. Meratnia, G. Heijenk, and H. Scholten, "A directional data dissemination protocol for vehicular environments," *Comput. Commun.*, vol. 34, no. 17, pp. 2057–2071, Nov. 2011.

[21] P.-C. Cheng, K. C. Lee, M. Gerla, and J. Härri, "Geodtn+nav: Geographic dtn routing with navigator prediction for urban vehicular environments," *Mobile Networks and Applications*, vol. 15, no. 1, pp. 61–82, 2010.

[22] S. A. Bitaghsir and F. Hendessi, "An intelligent routing protocol for delay tolerant networks using genetic algorithm," in *Smart Spaces and Next Generation Wired/Wireless Networking, 11th International Conference, NEW2AN 2011, and 4th Conference on Smart Spaces, Proceedings*, ser. Lecture Notes in Computer Science, vol. 6869. Springer, 2011, pp. 335–347.

[23] V. Soares, J. Rodrigues, and F. Farahmand, "Geospray: A geographic routing protocol for vehicular delay-tolerant networks," *Information Fusion (to appear)*, 2012.

[24] V. Naumov, R. Baumann, and T. Gross, "An evaluation of inter-vehicle ad hoc networks based on realistic vehicular traces," in *Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing*. ACM, 2006, pp. 108–119.

 ISBN: 978-1-61208-229-5