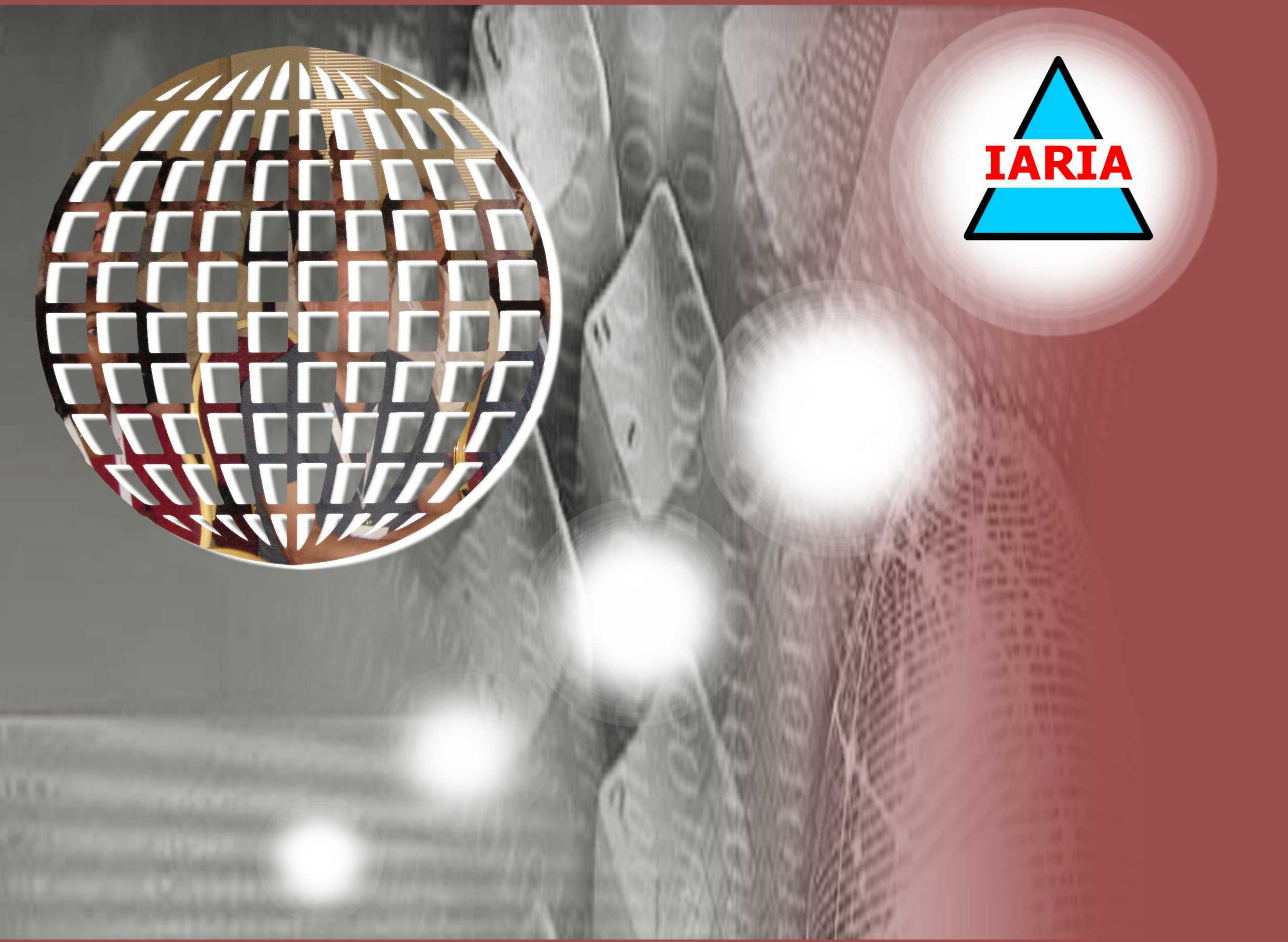
International Journal on
Advances in Telecommunications



2020 vol. 13 nr. 3&4

The *International Journal on Advances in Telecommunications* is published by IARIA.
ISSN: 1942-2601
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Telecommunications
Volume 13, Number 3 & 4, 2020

Philip L. Balcaen, University of British Columbia Okanagan - Kelowna, Canada
Marco Baldi, Università Politecnica delle Marche, Italy
Ilija Basicevic, University of Novi Sad, Serbia
Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Mark Bentum, University of Twente, The Netherlands
David Bernstein, Huawei Technologies, Ltd., USA
Eugen Borcoci, University "Politehnica"of Bucharest (UPB), Romania
Fernando Boronat Seguí, Universidad Politecnica de Valencia, Spain
Christos Bouras, University of Patras, Greece
Martin Brandl, Danube University Krems, Austria
Julien Broisin, IRIT, France
Dumitru Burdescu, University of Craiova, Romania
Andi Buzo, University "Politehnica" of Bucharest (UPB), Romania
Shkelzen Cakaj, Telecom of Kosovo / Prishtina University, Kosovo
Enzo Alberto Candreva, DEIS-University of Bologna, Italy
Rodrigo Capobianco Guido, São Paulo State University, Brazil
Hakima Chaouchi, Telecom SudParis, France
Silviu Ciochina, Universitatea Politehnica din Bucuresti, Romania
José Coimbra, Universidade do Algarve, Portugal
Hugo Coll Ferri, Polytechnic University of Valencia, Spain
Noel Crespi, Institut TELECOM SudParis-Evry, France
Leonardo Dagui de Oliveira, Escola Politécnica da Universidade de São Paulo, Brazil
Kevin Daimi, University of Detroit Mercy, USA
Gerard Damm, Alcatel-Lucent, USA
Francescantonio Della Rosa, Tampere University of Technology, Finland
Chérif Diallo, Consultant Sécurité des Systèmes d'Information, France
Klaus Drechsler, Fraunhofer Institute for Computer Graphics Research IGD, Germany
Jawad Drissi, Cameron University , USA
António Manuel Duarte Nogueira, University of Aveiro / Institute of Telecommunications, Portugal
Alban Duverdier, CNES (French Space Agency) Paris, France
Nicholas Evans, EURECOM, France
Fabrizio Falchi, ISTI - CNR, Italy
Mário F. S. Ferreira, University of Aveiro, Portugal
Bruno Filipe Marques, Polytechnic Institute of Viseu, Portugal
Robert Forster, Edgemount Solutions, USA
John-Austen Francisco, Rutgers, the State University of New Jersey, USA
Kaori Fujinami, Tokyo University of Agriculture and Technology, Japan
Shauneen Furlong , University of Ottawa, Canada / Liverpool John Moores University, UK
Emiliano Garcia-Palacios, ECIT Institute at Queens University Belfast - Belfast, UK
Ana-Belén García-Hernando, Universidad Politécnica de Madrid, Spain
Bezalel Gavish, Southern Methodist University, USA
Christos K. Georgiadis, University of Macedonia, Greece
Mariusz Glabowski, Poznan University of Technology, Poland
Katie Goeman, Hogeschool-Universiteit Brussel, Belgium
Hock Guan Goh, Universiti Tunku Abdul Rahman, Malaysia
Pedro Gonçalves, ESTGA - Universidade de Aveiro, Portugal
Valerie Gouet-Brunet, Conservatoire National des Arts et Métiers (CNAM), Paris
Christos Grecos, University of West of Scotland, UK
Stefanos Gritzalis, University of the Aegean, Greece
William I. Grosky, University of Michigan-Dearborn, USA
Vic Grout, Glyndwr University, UK
Xiang Gui, Massey University, New Zealand
Huaqun Guo, Institute for Infocomm Research, A*STAR, Singapore

Song Guo, University of Aizu, Japan
Kamal Harb, KFUPM, Saudi Arabia
Ching-Hsien (Robert) Hsu, Chung Hua University, Taiwan
Javier Ibanez-Guzman, Renault S.A., France
Lamiaa Fattouh Ibrahim, King Abdul Aziz University, Saudi Arabia
Theodoros Iliou, University of the Aegean, Greece
Mohsen Jahanshahi, Islamic Azad University, Iran
Antonio Jara, University of Murcia, Spain
Carlos Juiz, Universitat de les Illes Balears, Spain
Adrian Kacso, Universität Siegen, Germany
György Kálmán, Gjøvik University College, Norway
Eleni Kaplani, University of East Anglia-Norwich Research Park, UK
Behrouz Khoshnevis, University of Toronto, Canada
Ki Hong Kim, ETRI: Electronics and Telecommunications Research Institute, Korea
Atsushi Koike, Seikei University, Japan
Ousmane Kone, UPPA - University of Bordeaux, France
Dragana Krstic, University of Nis, Serbia
Archana Kumar, Delhi Institute of Technology & Management, Haryana, India
Romain Laborde, University Paul Sabatier (Toulouse III), France
Massimiliano Laddomada, Texas A&M University-Texarkana, USA
Wen-Hsing Lai, National Kaohsiung First University of Science and Technology, Taiwan
Zhihua Lai, Ranplan Wireless Network Design Ltd., UK
Jong-Hyouk Lee, INRIA, France
Wolfgang Leister, Norsk Regnesentral, Norway
Elizabeth I. Leonard, Naval Research Laboratory - Washington DC, USA
Richard Li, Huawei Technologies, USA
Jia-Chin Lin, National Central University, Taiwan
Chi (Harold) Liu, IBM Research - China, China
Diogo Lobato Acatauassu Nunes, Federal University of Pará, Brazil
Andreas Loeffler, Friedrich-Alexander-University of Erlangen-Nuremberg, Germany
Michael D. Logothetis, University of Patras, Greece
Renata Lopes Rosa, University of São Paulo, Brazil
Hongli Luo, Indiana University Purdue University Fort Wayne, USA
Christian Maciocco, Intel Corporation, USA
Dario Maggiorini, University of Milano, Italy
Maryam Tayefeh Mahmoudi, Research Institute for ICT, Iran
Krešimir Malarić, University of Zagreb, Croatia
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Herwig Mannaert, University of Antwerp, Belgium
Michael Massoth, University of Applied Sciences - Darmstadt, Germany
Adrian Matei, Orange Romania S.A, part of France Telecom Group, Romania
Natarajan Meghanathan, Jackson State University, USA
Emmanouel T. Michailidis, University of Piraeus, Greece
Ioannis D. Moscholios, University of Peloponnese, Greece
Djafar Mynbaev, City University of New York, USA
Pubudu N. Pathirana, Deakin University, Australia
Christopher Nguyen, Intel Corp., USA
Lim Nguyen, University of Nebraska-Lincoln, USA
Brian Niehöfer, TU Dortmund University, Germany
Serban Georgica Obreja, University Politehnica Bucharest, Romania
Peter Orosz, University of Debrecen, Hungary
Patrik Österberg, Mid Sweden University, Sweden
Harald Øverby, ITEM/NTNU, Norway

Tudor Palade, Technical University of Cluj-Napoca, Romania
Constantin Paleologu, University Politehnica of Bucharest, Romania
Stelios Papaharalabos, National Observatory of Athens, Greece
Gerard Parr, University of Ulster Coleraine, UK
Ling Pei, Finnish Geodetic Institute, Finland
Jun Peng, University of Texas - Pan American, USA
Cathryn Peoples, University of Ulster, UK
Dionysia Petraki, National Technical University of Athens, Greece
Dennis Pfisterer, University of Luebeck, Germany
Timothy Pham, Jet Propulsion Laboratory, California Institute of Technology, USA
Roger Pierre Fabris Hoefel, Federal University of Rio Grande do Sul (UFRGS), Brazil
Przemyslaw Pochec, University of New Brunswick, Canada
Anastasios Politis, Technological & Educational Institute of Serres, Greece
Adrian Popescu, Blekinge Institute of Technology, Sweden
Neeli R. Prasad, Aalborg University, Denmark
Dušan Radović, TES Electronic Solutions, Stuttgart, Germany
Victor Ramos, UAM Iztapalapa, Mexico
Gianluca Reali, Università degli Studi di Perugia, Italy
Eric Renault, Telecom SudParis, France
Leon Reznik, Rochester Institute of Technology, USA
Joel Rodrigues, Instituto de Telecomunicações / University of Beira Interior, Portugal
David Sánchez Rodríguez, University of Las Palmas de Gran Canaria (ULPGC), Spain
Panagiotis Sarigiannidis, University of Western Macedonia, Greece
Michael Sauer, Corning Incorporated, USA
Marialisa Scatà, University of Catania, Italy
Zary Segall, Chair Professor, Royal Institute of Technology, Sweden
Sergei Semenov, Broadcom, Finland
Dimitrios Serpanos, University of Patras and ISI/RC Athena, Greece
Adão Silva, University of Aveiro / Institute of Telecommunications, Portugal
Pushpendra Bahadur Singh, MindTree Ltd, India
Mariusz Skrocki, Orange Labs Poland / Telekomunikacja Polska S.A., Poland
Leonel Sousa, INESC-ID/IST, TU-Lisbon, Portugal
Cristian Stanciu, University Politehnica of Bucharest, Romania
Liana Stanescu, University of Craiova, Romania
Cosmin Stoica Spahiu, University of Craiova, Romania
Young-Joo Suh, POSTECH (Pohang University of Science and Technology), Korea
Hailong Sun, Beihang University, China
Jani Suomalainen, VTT Technical Research Centre of Finland, Finland
Fatma Tansu, Eastern Mediterranean University, Cyprus
Ioan Toma, STI Innsbruck/University Innsbruck, Austria
Božo Tomas, HT Mostar, Bosnia and Herzegovina
Piotr Tyczka, ITTI Sp. z o.o., Poland
John Vardakas, University of Patras, Greece
Andreas Veglis, Aristotle University of Thessaloniki, Greece
Luís Veiga, Instituto Superior Técnico / INESC-ID Lisboa, Portugal
Calin Vladeanu, "Politehnica" University of Bucharest, Romania
Benno Volk, ETH Zurich, Switzerland
Krzysztof Walczak, Poznan University of Economics, Poland
Krzysztof Walkowiak, Wroclaw University of Technology, Poland
Yang Wang, Georgia State University, USA
Yean-Fu Wen, National Taipei University, Taiwan, R.O.C.
Bernd E. Wolfinger, University of Hamburg, Germany
Riaan Wolhuter, Universiteit Stellenbosch University, South Africa

International Journal on Advances in Telecommunications
Volume 13, Numbers 3 & 4, 2020

CONTENTS

Construction of Secure Internal Network with Communication Classifying System Using Multiple Judgment Methods

Hirokazu Hasegawa

Information Security Office,
Nagoya University
Nagoya, Japan
Email: hasegawa@icts.nagoya-u.ac.jp

Yuya Sato

ZOZO Technologies, Inc.
Tokyo, Japan
Email: yuya.sato@zozo.com

Hiroki Takakura

Center for Cybersecurity
Research and Development,
National Institute of Informatics
Tokyo, Japan
Email: takakura@nii.ac.jp

Abstract—**Recent sophistication of cyber attacks targeting organizations such as companies, governments, and so on, have made the complete protection of our network very difficult. However, with the conventional measures including intrusion detection systems or firewalls, our network is not completely safe from intrusion because the dedicated malwares can slip through such measures. Thus, the separated network is one of the most effective countermeasures. In the separated network, an organization's internal network is divided into multiple segments, and fine access control among separated segments is conducted. To support a separated network construction, an automated ACL generation system has been previously proposed because the separated network is difficult to construct. However, this method focuses on the business continuity of the organization, and ACL will unconditionally permit the communication of a section where traffic is observed to maintain business continuity. Therefore, we have proposed a communication classifying system to judge the necessity of communication and its protocol by a two-step investigation. First, the system judges the consistency of the communication permitted by conventional systems. Second, if inconsistent communication is detected, the system judges the validity of the communication by checking the waiting state of its destination terminal. However, the system misjudges the necessity of communication in several conditions. In this paper, to resolve the misjudgment of the conventional communication classifying system, we improve it to conduct statistical analysis as a third investigation. In the experiment, the proposed system detected and terminated unintended communication between clients and servers. Thus, the proposed system outperformed the conventional communication classifying system.**

Keywords*—*Targeted Attacks; Network Separation; Access Control; Statistical Analysis.

I. Introduction

This paper is follow up of our previous paper "An Evaluation on Feasibility of a Communication Classifying System" already published in the proceedings of SECURWARE 2019 [1].

Recently, cyber attacks targeting organizations such as specific companies or governments have frequently occurred. Such attacks are called targeted attacks, and attackers have specific purposes, e.g., information theft and sabotage activities. In contrast to conventional indiscriminate attacks committed for the fun of a solo attacker, targeted attacks are conducted by multiple attackers belonging to well-funded crime groups for money making. In order to reach the goal of the attack, attackers use sophisticated methods and continue the attack persistently. For example, they carefully investigate the target and prepare dedicated malwares against the target including zero-day attacks.

Generally, organizations have applied several cybersecurity measures. For example, firewalls and intrusion detection systems are located on the border between the internet and the internal network to prevent intrusion by malwares. In many cases, such countermeasures use pattern matching technology with known malicious information and probably cannot detect unknown attacks such as zero-day attacks. Therefore, when sophisticated attacks slip through these countermeasures, we cannot prevent their invasion.

Therefore, the sophistication of cyber attacks has made it very difficult to protect our network from the intrusion of malwares completely. Against such a situation, recent countermeasures have been focusing on after the intrusion of malwares. The goal of such countermeasures is the mitigation of damages by the attacks, e.g., preventing information leakage and ceasing file destruction activities [2].

A separated network is one of the effective countermeasures [3], and our research has been focusing on it. It divides the organization's internal network into multiple segments and performs fine access control among the divided segments. In the conventional network structure, only a single segment without any access control is deployed, that is, all terminals are connected to the segment, and they can directly communicate with all others. In contrast to such a traditional structure, the separated network restricts communication in the internal network, and we can prevent unintended communication caused by malwares, e.g., lateral movement. In addition, when we detect malwares, it can minimize the harmful effect on business continuity because we can isolate only the infected segment or terminals.

It is difficult to construct and maintain the separated network because the border among segments and its access controls must be determined using various information concerning networks, human resources, business contents, and so on. Moreover, the change in human resources or business contents should be followed to maintain access control. Therefore, the separated network is not cost-effective, and many organizations still use traditional structures in the internal network.

In our earlier work, we proposed several systems to solve such problems and to support constructing a separated network. In our research, we assume a general organization's network is divided into several segments based on the department. Our goal is to construct a separated network by applying fine access controls permitting only necessary communication to network equipment that is the border of each segment. A necessary communication is defined as legitimate communication required for the works of users. For example, when a user needs to access the file in server A, the communication between the user and server A is defined as necessary. On the other hand, when a user never uses server B, the communication between the user and server B is defined as unnecessary.

An automated ACL generation system is our first work [4], and the system generates ACL automatically based on the user's access authority against files or directories in the servers and on existing communication in the network. We call this system "AAGS (Automated ACL Generation System)" in this paper. Although it reduces the burden of separated network construction by the administrators, the generated ACL allows overly permission and prohibition of the communication.

AAGS may permit several unnecessary communications. To avoid such overly permission, we need a detailed judgment method for the necessity of communication. Thus, we proposed a communication classifying system [5] to avoid overly permission. In this paper, we call the system a Communication Classifying System "(CCS)". The system carefully investigates existing communication and evaluates the consistency and the validity of the observed communication by checking the state of its destination terminal.

Here, we improved and implemented the CCS, and we verified its feasibility in our experimental network [1]. As a result of the experiment, we found several problems. These problems make CCS misjudge the necessity of communication in several cases.

Moreover, we improve the CCS to conduct a three-step investigation for verification of communication necessity. We deploy a statistical analysis for the third investigation to solve the problem of the CCS. Therefore, we implemented an extended CCS and experiment to evaluate the feasibility of the system.

The rest of this paper is organized as follows: In Section II, we introduce researches against targeted attacks and related works. Section III presents the proposal system. The implementation of the proposed system explained in Section IV. In Section V, we describe the evaluation of the proposal system. Finally, we conclude this paper in Section VI.

II. Related Works

In this section, we introduce related works for mitigating targeted cyber attacks and our previous research.

A. Research for Preventing Malware Activities

Many researchers have done several works to prevent+ malware activities in internal networks. Alessandro et al. proposed a method for modeling communication patterns of malwares that perform lateral movement [6]. However, it is not cost-effective to employ this method because we need to install a communication analysis tool on all terminals in the network. In the separated network, the spread of infection can be suppressed without installing special tools on the terminal because the ACL limits the communication area of malwares.

Also, several methods to construct the separated network have been widely studied. Watanabe et al. proposed a VLAN (Virtual Local Area Network) configuration method [7]. This method monitors traffic in the network and generates a network design using the monitoring information. When a certain amount of traffic exceeding the threshold among terminals is observed, a new VLAN concerning the terminals is generated. Because the terminals are frequently communicating with each other, it is effective from the viewpoint of the amount of traffic volume. However, if the VLAN is generated, the infected terminal is in the VLAN, it cannot prevent malware activities in that VLAN. According to [8] [9] [10] in supporting VLAN construction, the works do not pay attention to constructing moderate access controls among VLANs because such works only focus on the network efficiency. Besides the above researches, several products, e.g., "VLAN.Config" [11], for constructing VLAN automatically are difficult to generate ACL for the constructed network.

On the other hand, some researchers focus on the segmentation of the internal network for security measures. Mujib et al. constructed a micro-segmentation environment by using Cisco Application Centric Infrastructure (ACI) and evaluated the effectiveness of micro-segmentation against cyber attacks [12]. The result shows that the micro-segmentation is effective against cyber attacks. However, it is not shown what criteria should be used to construct micro-segmentation in a real environment network. Wagner et al. proposed a semi-automated network segmentation construction method in [13]. Moreover, in [14], they proposed a fully-automated network segmentation generation method focusing on security, cost, mission performance. Their proposals are effective methods for constructing a segmented network, however, their methods are based on simulation using network environmental data and attack threat data. The preparation of such data is not cost-effective. Our method is more effective from the viewpoint of cost and user convenience because it is based on the real network traffic data and coordinates the access control to the user's real traffic.

From the viewpoint of traffic investigation of the internal network, there are many researches for malware activity detection [15] [16]. However, recently the encryption of communication is often conducted, and malwares also encrypt their communication for avoiding detection. There are many researches for decryption of communication for detecting malware communication [17], however, it includes the problem of privacy. In our research, the proposed system can treat observed traffic regardless it is encrypted or not.

B. Automated ACL Generation System (AAGS)

This research focuses on the separated network, and we proposed several systems that support the separated network construction. An AAGS [4] evaluates the necessity of communication sections based on two criteria, i.e., access authority of a user to files or directories in servers and existing communication in an internal network.

Generally, the access authority of files or directories is strictly managed. For example, although all members can access public information, confidential information is managed for access by the only concerned person(s). Thus, the communication section between a user and a server is unnecessary if the user has no access authority to all files and directories in the server. Many organizations apply directory service for managing access authority, therefore, the system gathers information on access authorities by analyzing the information in the directory server in a network and evaluates the necessity of communication sections. By using the result of the evaluation, AAGS generates ACL automatically.

However, there are various types of necessary communication in a network except for file access communication. The system judges such legitimate communication as unnecessary if it evaluates based only on file access authority. To avoid such a situation and secure business continuity, AAGS analyzes the mirrored packets of the internal network. Before applying the ACL generated based on file access authorities, the system revises it by using mirrored packets. Even if communication was previously judged as unnecessary, its new observation calls reevaluation, and the communication is judged as necessary. Based on the judgment, the system regenerates ACL to permit all of the necessary communication sections. Thus, AAGS supports us to construct the separated network easily by applying the generated ACL.

C. Problems of the AAGS

Although AAGS can reduce the burden of administrators in constructing or managing the separated network, the generated ACL is not properly described because of the following two reasons.

First, AAGS judges all communications observed in the network as necessary even if they occur unintentionally, and it permits all of such communications. Therefore, generated ACL may include overly permission of unnecessary communication sections.

Second, the ACL only is based on the pair of source and destination IP addresses. Once the system judged the communication section to be allowed, all communication protocols on the pair are permitted.

III. Communication Classifying System (CCS)

To solve the problems of AAGS, we proposed a CCS [5] that improves the preciseness of ACL generated by AAGS.

A. System Overview

Previously, we made CCS conduct a two-step investigation. First, CCS investigates the consistency between the communication observed in the network and the reason AAGS permitted such communication section, i.e., user's file access authority or communication observation. When the observed communication does not relate to file sharing even though the file access authority is the reason permitting the communication section, such communication does not have consistency. If a communication lacks consistency, CCS performs the additional investigation. Because a legitimate communication assumes that its destination terminals listen to the appropriate port, the system performs a port scan to identify the listening port and then analyzes the correlation between the observed communication protocols and listening ports of destination terminals.

These investigations make CCS possible to detect illegal communication. Finally, the system generates a new ACL described by the sets of source and destination IP addresses and destination port to permit only legitimate communication and prohibiting unnecessary communication.

However, we noticed that the CCS makes misjudgments in several conditions. For example, the previous experiment made a client conduct illegal SMB communication to a file server. Although the client has no access authority to the server, CCS judged such communication as necessary. CCS judged such communication as necessary. Such misjudgments occur in the condition that a destination server provides the same service as illegal communication for specific users. For example, as shown in Figure 1, we assume a case that the server provides web services against only regular staff excluding part-time staff by using the authentication function. In this time, the server listens 80/tcp port for using the HTTP protocol. If a part-time staff accidentally communicates with the server via the HTTP protocol, the web service cannot be used because this server cannot be authenticated. In other words, this accidentally occurred communication is an unnecessary one. However, CCS verified such communication as a necessary one as the destination server opens the corresponding service port of the communication.

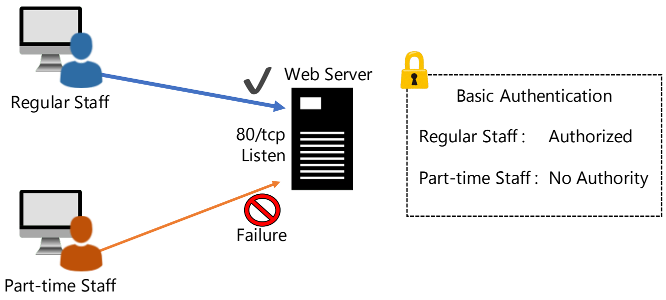


Fig. 1. Example of a condition in which judgment fails.

Moreover, this paper extends the CCS to conduct a three-step investigation by combining the conventional methods and the new statistical judgment method.

B. Assumption in CCS

We proposed CCS to complement our previous AAGS. CCS assumes that the network is roughly divided into several segments, and ACL generated by AAGS is applied to the network. The applied ACL is stored in a database (ACL DB) by AAGS.

ACL DB is extended by adding three new columns. First, we added "Permitted Reason" to register the reason why the communication is permitted, i.e., directory service information, or/and communication analysis. AAGS uses the extended versions of DB so that the ACL describes permitted communication sections, e.g., source IP addresses, destination IP addresses, and Permitted Reason. The remaining two columns are "Destination Port" and "Status". However, AAGS ignores these two columns as empty fields.

When CCS analyzes the communication section, it registers "analyzed" to the Status field of such a communication section. If there is only one record for the pair of source IP address and destination IP address, and such a record's Status field is empty, it is the first time for the proposed system to analyze that communication section. If "analyzed" has been registered to the Status field of a communication section, CCS omits the analysis of the communication section.

In addition to the above case where AAGS or CCS permitted the communication section, we assume the other case that different ways permit the communication section. For example, an administrator can permit any protocols manually. Furthermore, the research "Dynamic Access Control System" [18] associates with CCS to permit communication overly prohibited by AAGS. In these cases, the permitted communication section has not yet been analyzed by CCS. To distinguish the not analyzed communication, CCS assumes that "not_analyzed" is registered to Status filed of the communication section not permitted by AAGS or CCS. If the Status field is "not_analyzed", CCS analyzes the communication protocols in the section.

In this paper, to simplify the discussion, we assume that all terminals are statically assigned IP addresses, and such assignment information is managed in a directory server. However, our method can be easily applied to the environments that employ dynamically IP address assignment methods, e.g., DHCP. We can control the connected device's communication by identifying the device's user with any authentication method, e.g., IEEE 802.1X. For example, we can assign the appropriate VLAN that the user should belong to, or update ACL based on the assigned IP address.

C. The architecture of CCS

Figure 2 shows the architecture of CCS. The system consists of six modules and the extended database in the AAGS. The details of each module are described below.

1) Traffic Collector: This module receives all mirrored packets generated in the internal network. This paper assumes that the collection period of mirrored packets for investigation is statically defined in advance, e.g., 1 day, 1 hour, 10 minutes, and so on. After mirrored packets collection, the module generates a list of packet information including sets of source IP address, destination IP address, and destination port from the collected packet. The generated list of packet information is sent to the Consistency Judgment module.

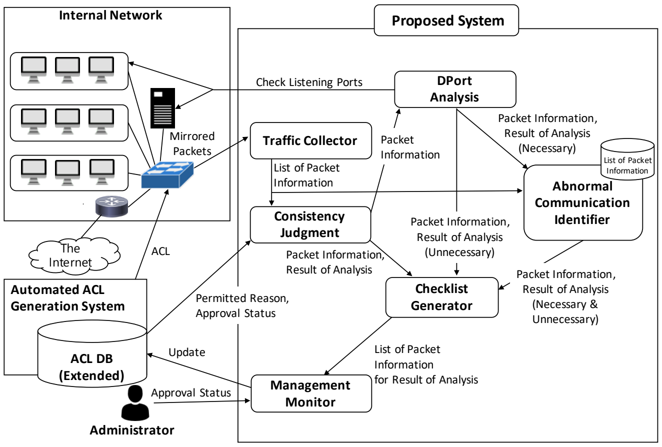

Internal Network
Proposed System
Check Listening Ports
DPort Analysis
Mirrored Packets
Traffic Collector
Packet Information
Packet Information, Result of Analysis (Necessary)
List of Packet Information
List of Packet Information
Abnormal Communication Identifier
Consistency Judgment
Packet Information, Result of Analysis (Unnecessary)
The Internet
ACL
Packet Information, Result of Analysis
Packet Information, Result of Analysis (Necessary & Unnecessary)
Automated ACL Generation System
Permitted Reason, Approval Status
Checklist Generator
ACL DB (Extended)
Update
List of Packet Information for Result of Analysis
Approval Status
Management Monitor
Administrator

Fig. 2. The architecture of a communication classifying system.

2) Consistency Judgment: First, when a list of packet information is received, this module searches ACL DB records for each communication section by specifying each pair of source and destination IP addresses. When the status field is empty, the Consistency Judgment module analyzes all protocols captured in that communication section.

After extracting the subject of the communication for investigation, the Consistency Judgment module judges the consistency of such communication. The module finds the permitted reason for such communication by checking ACL DB. As shown in Table I, there are six combinations of a collected packet and communication reasons. In the table, CA denotes communication analysis. Because AAGS checks the necessity of the file-sharing communication by using a Directory Service Information (DSI), the Consistency Judgment module classifies the captured communication as SMB protocol or Other Protocols. In this paper, we assume that only SMB is used as a file-sharing communication protocol. SMB uses multiple ports and protocols, e.g., 139/tcp and 445/tcp. To simplify the discussion, we express these sets of all ports by using the term "SMB protocol".

TABLE I
COMBINATIONS OF PERMITTED REASON AND COLLECTED PACKET.

Collected Packet	Permitted Reason		
	DSI	DSI+CA	CA
SMB	1	2	3
Other Protocol	4	5	6

For the SMB protocol, combinations 1 and 2 of Table I have consistency. To permit these communications, the Consistency Judgment module sends this Packet Information to the Check List Generator module. On the other hand, communication lacks consistency in combination 3 because communication of SMB protocol was observed even if there was no access authority by DSI. However, file sharing may be conducted among user's terminals directly without management by the directory server. In order not to prohibit such communication, the Consistency Judgment module sends this Packet Information to the DPort Analysis module for further verification.

Apart from SMB, only combination 4 lacks consistency in other protocols. The Packet Information of such communication is sent to the Checklist Generator module to prohibit such communication. Though combinations 5 and 6 have consistency, the module cannot determine the sameness of the communication protocol collected by the Traffic Collector and AAGS. The Packet Information of such communication is sent to the DPort Analysis module, which conducts a detailed investigation.

3) DPort Analysis: This module analyzes the normality of communication. We assume that the destination terminal listens to the correct port of service for communication. In such an assumption, the module judges the normality of communication using the current stand-by states of destination terminals. There are several ways to specify the listening ports of terminals. Thus, we adopt port-scanning against destination terminals in this paper.

Based on the result of port-scanning, when the destination port of communication listened to the destination terminal, the DPort Analysis verifies that communication is necessary. When the judgment result is necessary, Packet Information and the result are sent to the Abnormal Communication Identifier module.

On the other hand, if the destination port is blocked, the communication is judged as unnecessary. Thus, the judgment result is sent to the Checklist Generator module with its packet information.

4) Abnormal Communication Identifier: Abnormal Communication Identifier module finds out abnormal communication from the Packet Information judged as necessary communication by the DPort Analysis module. In this paper, we define abnormal communication as the unintentional user's communication against a server that provides service for limited users.

To find the abnormal communication, we use statistical analysis. The module receives a list of packet information from the Traffic Collector module and stores it. Against the stored information, the module statistical analysis and finds abnormal communication.

As a basic idea, we assume that abnormal communication is extremely less than legitimate communication. For example, as shown in Figure 3, there is a web server that is utilized for managers. Here, communication via HTTP protocol is allowed between the server and the managers. On the other hand, communication accidentally conducted by part-time staff is very few compared to legitimate communication. The difference in the amount of communication between legitimate and anomalous communication can make a significant difference. Therefore, we intend to judge such very few volume communications as abnormal communication.

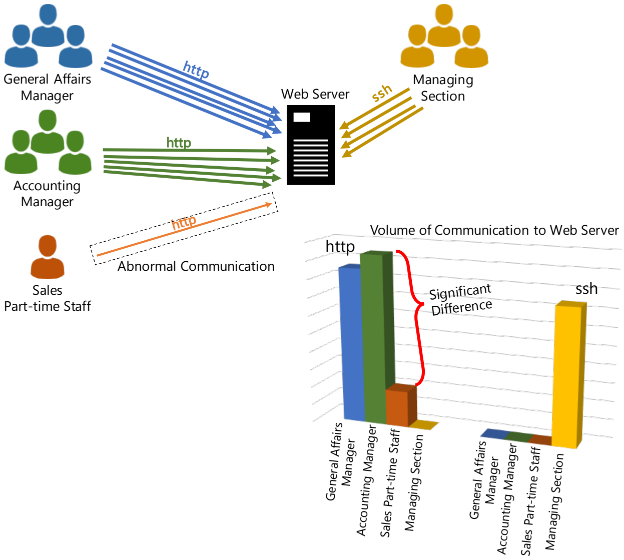


Fig. 3. The basic assumption of statistical analysis.

However, such a tendency of the communication volume depends on the communication protocol. When the web server shown in Figure 3 provides SSH for the system managing section, a lot of SSH, different from HTTP communication, is conducted only between the web server and the managing section.

From the viewpoint of this idea, we classify the communication based on the destination IP address and port number. When the Abnormal Communication Identifier receives the packet information to be verified from the DPort Analysis module, it extracts all the packet information with the same destination IP address and destination port number as the packet information to be verified. The module also counts the number of packets for each source IP address from all extracted packet information. We utilize these counted number of packets as data for statistical analysis.

Because various communication is allowed in the organization's network, and traffic volume depends on the communication content, we assumed that the data might contain an extremely large or small value. For such data, it is not appropriate to use non-robust statistics such as mean and standard deviation. In this paper, we adopt an outlier test using a quartile and an interquartile range (IQR) to consider robustness because a quartile and IQR are less affected by the size of the data values.

A quartile is a quantile dividing the data sorted in ascending order of value into four equal parts. The second quartile (Q_2) is the median, and it divides data into two equal parts. The first quartile (Q_1) is the median of data smaller than Q_2. It divides whole data into the lowest 25%. The third quartile (Q_3) is the median of data bigger than Q_2. It divides whole data into the lowest 75%. By using Q_1 and Q_3, we obtain IQR.

$$\mathrm{IQR} = Q_3 - Q_1 \tag{1}$$

By using a quartile and IQR, we set a threshold for the outlier test. In this paper, we apply a report of the tabulating statistical survey [19]. In [19], Noro and Wada pointed out that we can properly detect the outlier by using a threshold based on quartile and IQR even when the data does not follow a normal distribution. They used equation (2) for setting a lower limit of an appropriate range of data.

$$\mathrm{L} = Q_1 - 1.724 \times \mathrm{IQR} \tag{2}$$

However, the number "1.724" in the equation can be arbitrarily set according to the length of the tail of the distribution of the actual data. For simplicity of discussion, this section follows [19] and uses "1.724" without change.

In the communication data including extremely large or small values, IQR becomes so large that equation (2) cannot derive the threshold correctly. To solve such a problem, we convert all data in logarithm with base e in advance.

Because a threshold that shows a negative value cannot be properly treated, it is replaced with a positive value using an exponential conversion. Therefore, we earn a threshold by using equation (3) shown below.

$$\text{Threshold} = \begin{cases} L & (L >= 0) \\ e^L & (otherwise) \end{cases} \tag{3}$$

The Abnormal Communication Identifier module determines the communication is abnormal if the count is less than the threshold. If the abnormal communication is detected, the judgment result of this communication is changed as unnecessary. Finally, the results and Packet Information are sent to the Checklist Generator module.

5) Checklist Generator: This module receives the packet information and judgment results from the Consistency Judgment module or DPort Analysis module. The Checklist Generator module combines the packet information and analysis results and generates a checklist from this information for administrators. The generated checklist of the packet information is sent to the Management Monitor module.

6) Management Monitor: Lists of the packet information and judgment results are sent from the Checklist Generator module to the Management Monitor module. This module presents the received lists to the administrators. Administrators check the list and authorize the permission or prohibition of the communication section. Finally, the module updates the ACL DB to register the authorized packet information as "analyzed" value in the status field. After updating the ACL DB, the ACL Applier in AAGS applies it to the network.

IV. Implementation of the Proposed System

This section describes the implementation of the proposed system. The basic structure of the modules and the data flow among modules are shown in Figure 4. In the proposed system, the Traffic Collector module, the Consistency Judgment module, the DPort Analysis module, and the Abnormal Communication Identifier module run as batch processing written with Python. The list of observed packet information, in which the Abnormal Communication Identifier stores designed using MySQL database [20]. We adopt Node.js [21] as a Web server including the Checklist Generator module and the Management Monitor. Also, we designed an API server by using FastAPI [22] for smooth data exchanges between each module and the ACL DB or between the Abnormal Communication Identifier module and the list of packet information.

In this paper, we implemented ACL DB and ACL Applier that are included in AAGS. In addition to the list of packet information, we used MySQL for ACL DB. By using the Software Designed Networking (SDN) technique, we realized the ACL applier. We assume that Open vSwitch [23] (OvS) is used as a network switch, and we adopted Trema [24] as the SDN controller that instructs the OvS to control packets in the network.

Moreover, all of these modules run on Docker [25], which manages applications using a container type virtual environment.

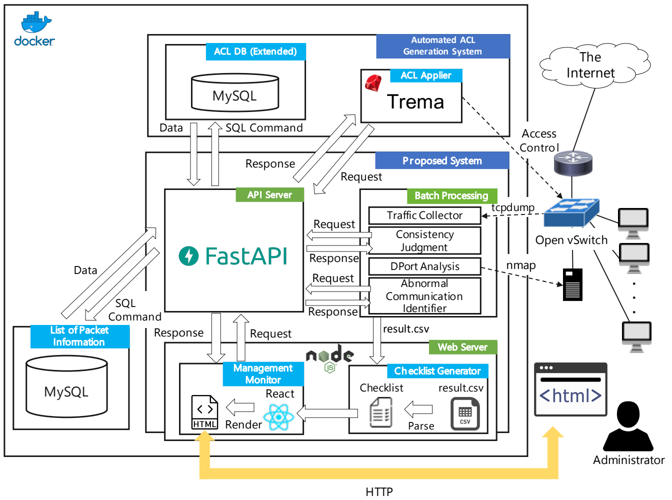


Fig. 4. System configuration diagram.

A. Traffic Collector

This module receives mirrored packets from the OvS and generates a list of packet information. We configure the OvS in advance to generate mirrors of all packets in the network and send them to the Traffic Collector. The Traffic Collector executes the `tcpdump` command and captures all of the mirrored packets sent from OvS for a collection period. In the experiment mentioned in Section V, we set a collection period as 1 hour or 30 minutes or 10 minutes.

The captured packets are saved as pcap files, and this module extracted sets of source IP address, destination IP address, and destination port for each packet from the saved pcap file by using dpkt [26], which is a module of Python. Finally, this module sends the extracted set as packet information to the Consistency Judgment module and Abnormal Communication Identifier module.

B. Consistency Judgment

After receiving the list of packet information, this module sends a request to the API server to search the record of the communication section in the ACL DB corresponding to each packet information. This module also checks the destination ports of each packet information and classifies them into SMB or other ports.

Further, the module compares such destination ports and the result of the record search, and it judges the consistency of the communication. When the module decides whether the observed communication is necessary or not, it sends the packet information of that communication section with the judgment results to the Checklist Generator module. On the other hand, if the module determines that the detailed analysis is necessary, it sends the packet information to the DPort Analysis module.

C. DPort Analysis

This module judges the normality of the communication included in packet information sent from the Consistency Judgment module. To assess the listening ports of destination terminals, it uses the `nmap` command. Here, we use the `-S` optional command of nmap to spoof the source IP address of the observed communication.

Based on the results of nmap, if the proper service port of packet information is listening at the destination terminal, the module judges this communication as rightful and necessary. Otherwise, the communication is judged as unnecessary.

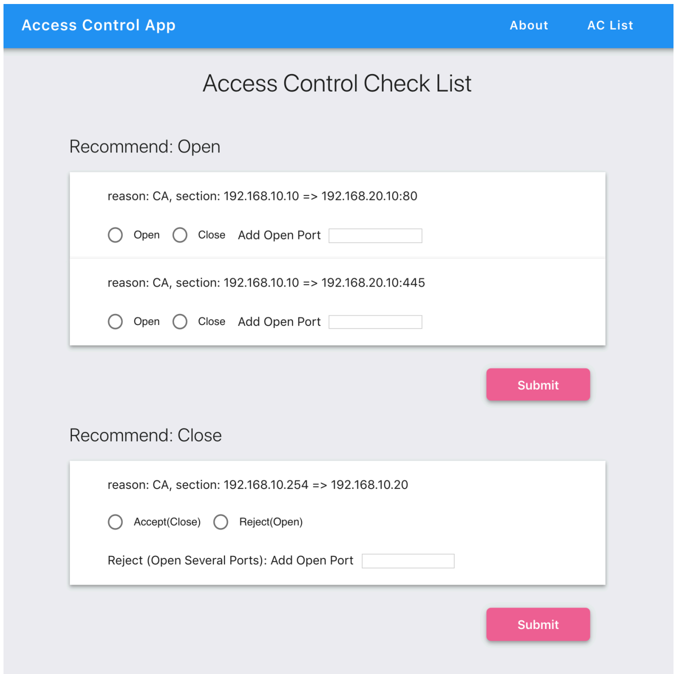
Moreover, if DPort Analysis determines the communication as unnecessary, the module sends the packet information and its judgment results to the Checklist Generator module, otherwise, communication is judged as necessary. Further, the packet information and its judgment results are sent to the Abnormal Communication Identifier module.

D. Abnormal Communication Identifier

This module receives all the list of packet information from the Traffic Collector module, and the module sends a request to the API server to store received packet information in the database. When the packet information and the judgment result are sent from the DPort Analysis module, the Abnormal Communication Identifier module sends a request to the API server to search the packets that have the same destination IP address and the same destination port of received packet information. By using receiving packet information and packet information found in the list of the packet information database, the module conducts the statistical analysis. Finally, the module sends packet information and judgment result to the Checklist Generator module.

E. Checklist Generator and Management Monitor

The Checklist Generator module receives the packet information and its judgment results from the Consistency Judgment module and the DPort Analysis module. The Checklist Generator combines these pieces of information about the packet and generates the checklist of the packet information.



Access Control Check List

Recommend: Open

reason: CA, section: 192.168.10.10 => 192.168.20.10:80

○ Open ○ Close Add Open Port

reason: CA, section: 192.168.10.10 => 192.168.20.10:445

○ Open ○ Close Add Open Port

Recommend: Close

reason: CA, section: 192.168.10.254 => 192.168.10.20

○ Accept(Close) ○ Reject(Open)

Reject (Open Several Ports): Add Open Port

Fig. 5. Sample of management monitor web page.

The generated list of packet information is sent to the Management Monitor. Then, based on the list, a html page is generated as an interface for administrators by using React [27]. Figure 5 shows a sample of the generated Web page for administrators.

On this screen, there are two sections. The first section is “Recommend: Open”. The communication sections displayed in this section is judged as necessary. If the administrator judges it as appropriate, it can be authorized by selecting the “Open” button. However, only the displayed ports are judged necessary by the system, and all of the other ports not displayed will be prohibited. When administrators want to permit several ports in addition to the system recommendation, they can insert such ports into the “Add Open Port” form. Otherwise, they use the “Close” button to prohibit the displayed communication.

The second section is “Recommend: Close”. The system judged communication displayed in this section as unnecessary. If the administrator selects the “Accept (Close)” button, all communication in this section is prohibited. On the other hand, when the “Reject (Open)” is selected, the ACL permits all communication in this section. Also, if the administrator wants to permit several ports in this section, such ports can be inserted into the “Add Open Port” form.

Finally, this module updates the ACL DB using the API server when the “Submit” button is clicked. As mentioned in the next Subsection IV-F, the ACL DB stores only permitted communication sections. In case of that all analyzed communication is judged as still permitted, the system updated the status field of the flow_list table about such communication section as analyzed. If only several ports are permitted, in addition to the above update, those ports are inserted into the dst_port field.

On the other hand, if all protocols in the communication section are judged as unnecessary, the module updates the ACL DB to delete any record of such a communication section in the section_list table.

F. ACL DB (Extended)

As described in Section III-C, we extended ACL DB. ACL DB consists of two tables, "section_list" and "flow_list" shown in Table II. The section_list table consists of four columns: "id", "src_ip", "dst_ip", and reason. The src_ip and the dst_ip store the source IP address and destination IP address of the communication section permitted by AAGS respectively. The reason column stores the permitted reason.

TABLE II
ACL DB (EXTENDED) TABLE SCHEMA.

Table Name	Column	Data Type	Example
section_list	id	Integer	3
	src_ip	String	192.168.10.10
	dst_ip	String	192.168.20.20
	reason	String	CA
flow_list	section_id	Integer	3
	dst_port	Integer	443
	status	String	analyzed

The flow_list table consists of three columns: "section_id", "dst_port", and "status". The value of section_id is corresponding to the id of the section_list table. Permitted destination ports in the communication section are stored in the dst_port column. If the communication section is permitted without analysis by CCS, "not_analyzed" is stored in the status column. After analysis by CCS, the value of status is updated to "analyzed".

G. ACL Applier

We use the SDN technique to implement the ACL Applier. The OvS is operating as a core switch in the network. We use Trema as an OpenFlow controller to apply the contents of ACL DB to the network.

V. EVALUATION EXPERIMENT

We applied the implemented system to our prototype network and conducted an evaluation experiment on it.

A. Experimental Conditions

1) Network Structure: Here, we conducted experiments in our prototype network to verify the various situations by extending our prototype network, which is used for evaluation of the CCS [1]. As shown in Figure 6, the network has one sever segment and four client segments.

Each segment has seven or four Windows 10 PCs, and all PCs are assigned static IP addresses. Besides, two Windows Server 2019 terminals are located in the server segment. One of these servers works as a file server, and another server works as an active directory, which also has the role of DNS in this organization. The file server permits access only from the user whose position is the manager.

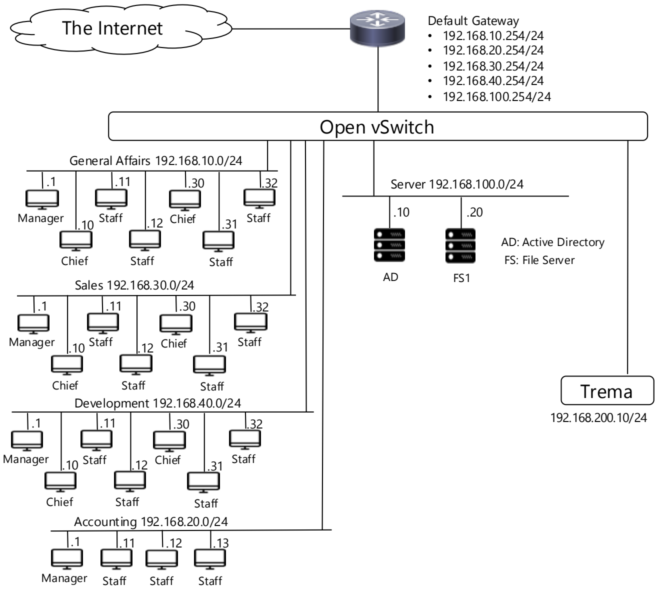


Fig. 6. Prototype network architecture.

2) Access Controls: The AAGS generated the ACL, and we prepared the ACL shown in Table III. We configured Trema to permit only the communication listed in Table III in addition to the communications between the default gateway and all the terminals. Open vSwitch, controlled by Trema, performs the access control.

TABLE III
LIST OF COMMUNICATION SECTIONS PERMITTED BY AAGS.

Source IP Address	Destination IP Address	Permitted Reason
192.168.10.1	192.168.100.20	DSI
192.168.20.1	192.168.100.20	DSI+CA
192.168.20.11	192.168.100.20	CA
192.168.30.1	192.168.100.20	DSI+CA
192.168.40.1	192.168.100.20	DSI

Because the managers have access authority to the file server, 192.168.10.1, 192.168.20.1, 192.168.30.1, and 192.168.40.1 are permitted to communicate with 192.168.100.20, and we insert "DSI" as Permitted Reason in the records of these communication sections.

In addition to the file access communication, unintended communication from 192.168.20.1 and 192.168.30.1 to 192.168.100.20 are conducted, and "CA" is added to Permitted Reason of that section. Similarly, communication between 192.168.20.11 and 192.168.100.20 is permitted because of unintended communication, and "CA" is registered as its Permitted Reason.

B. Experiment 1: Judgments of All Communication

To evaluate the effectiveness of CCS, we run CCS to collect and judge all communication in the prototype network. The experiment was performed according to the following procedure.

Step 1: Run the proposed system and start to collect mirrored packets in the network. In this experiment, we set the collection period to be 1 hour.

Step 2: In the collection period, terminals, 192.168.10.1, 192.168.20.1, 192.168.30.1, and 192.168.40.1, access the file server using the SMB protocol. Also, the terminal of 192.168.20.11 that has no access authority tried SMB protocol communication with the file server. Though the file server does not provide HTTP service, HTTP protocol communication to the file server is conducted by terminals 192.168.20.1 and 192.168.30.1. In addition, all nine client terminals access external sites on the Internet that are assuming the activities of the organization.

Step 3: After 1 hour, the collection period ends, and the captured packets are analyzed by the proposed system. Based on the analysis result, the system generates the checklist and prepares the Web page.

Step 4: We check the result of the analysis by the proposed system on the Web page and authorize them.

Step 5: Finally, the system applies the authorized ACL to the internal network.

TABLE IV
ANALYSIS RESULT BY OUR PROPOSED SYSTEM.

Internal Network Communication that Occurred			Result of Analysis
Source IP Address	Destination IP Address	Destination Port	
192.168.10.1	192.168.100.20	445	Open
192.168.20.1	192.168.100.20	445	Open
192.168.20.11	192.168.100.20	445	Outlier
192.168.30.1	192.168.100.20	445	Open
192.168.40.1	192.168.100.20	445	Open
192.168.10.1	192.168.100.10	53	Open
192.168.10.10	192.168.100.10	53	Open
192.168.20.1	192.168.100.10	53	Open
∼	∼	53	Open
192.168.20.1	192.168.100.20	80	Close
192.168.30.1	192.168.100.20	80	Close
192.168.100.20	192.168.10.1	63221	Close
192.168.100.20	192.168.20.1	59012	Close
192.168.100.20	192.168.20.11	55658	Close
192.168.100.20	192.168.30.1	52796	Close
192.168.100.20	192.168.40.1	51166	Close
192.168.100.10	192.168.10.1	63205	Close
192.168.100.10	192.168.10.10	65180	Close
192.168.100.10	192.168.10.11	61426	Close
∼	∼	∼	Close

C. Results of Experiment

The result of the analysis using the proposed system is shown in Table IV. The legitimate SMB communication from 192.168.10.1, 192.168.20.1, 192.168.30.1, and 192.168.40.1 to the file server (192.168.100.20) is correctly judged as necessary. Also, the system judges the DNS protocol communication as necessary. Moreover, the system judges several communication sections including the unintended HTTP communication and high port number communication that seems to be returned packets as unnecessary.

Thus, the above results are approximately the same as the previous CCS's results. In this evaluation, we focus on communication from 192.168.20.11 to 192.168.100.20. The previous CCS judges the communication as necessary because the destination port has listened. On the other hand, the extended CCS judged such illegal communication as unnecessary as the "Outlier" is shown in the result of the analysis.

D. Experiment 2: Using Several Patterns of SMB Communication

In the experiment using all communication, the proposed system found abnormal communication correctly. To verify the credibility of the statistical judgment method, we further conducted another experiment. In this experiment, as same as experiment 1, four legitimate client terminals and one illegal client terminal tried to communicate with the file server. However, we conducted several patterns of experiments with different access counts or observation times. We only focus on these file-sharing communications and show the detailed process of the statistical analysis.

We generated different four access patterns shown in Table V. First, in pattern 1, legitimate clients frequently communicate with the file server, and the illegal client also conducts communication most frequently. We set the observation time as 30 minutes. In pattern 2, legitimate clients conduct communication as same as pattern 1. In contrast to pattern 1, the illegal client tried to communicate only once in 30 minutes. We set different observation time in pattern 3. Similar to pattern 2, the illegal client tried to communicate only once in 10 minutes. Finally, in pattern 4, all terminals randomly communicate with the file server at the same time. However, in all patterns, all terminals share different files with the file server. Even if all file-sharing communication is conducted at the same time, traffic volumes of each communication are different because of the file size.

In Table VI, the upper rows of each pattern show the number of observed packets, and the data for the statistical analysis that is generated by converting the number of observed packets in logarithm with base e is shown in the lower row.

Table VII shows the results of the statistical analysis. In patterns 1, 2, and 4, illegal communication's data for statistical analysis is judged as outlier correctly. On the other hand, only in pattern 3, the illegal communication's data (4.190) exceeds the threshold (3.221), and no outlier value was detected.

E. Discussion

From the results of experiment 1 and experiment 2, we found that the proposed system correctly judged the illegal SMB communication from 192.168.20.11 to 192.168.100.20 as unnecessary. Therefore, the judgment accuracy of the extended CCS is improved compared to the previous CCS that previously judged the communication as necessary.

In the case of pattern 1 in experiment 2, we expected that to conduct judgment correctly might be difficult because the illegal terminal generated communication most frequently in all terminals. However, correct judgment was conducted by the proposed system. As shown in Table VI, the observed number of packets for each terminal is different. The variation in the number of packets occurred because the traffic volume depends on the size of sharing files. Also, only traffic of protocol negotiation occurs in illegal communication. In other words,

TABLE V
ACCESS PATTERNS.

	192.168.10.1	192.168.20.1	192.168.30.1	192.168.40.1	192.168.20.11
Pattern 1 (30m)	Once / 20s	Once / 30s	Once / 15s	Once / 25s	Once / 5s
Pattern 2 (30m)	Once / 20s	Once / 30s	Once / 15s	Once / 25s	Once / 30m
Pattern 3 (10m)	Once / 20s	Once / 6s	Once / 7s	Once / 27s	Once / 10m
Pattern 4 (30m)	Random (All terminals communicate at the same time)				

TABLE VI
NUMBER OF OBSERVED PACKETS AND DATA USED FOR STATISTICAL ANALYSIS.

	192.168.10.1	192.168.20.1	192.168.30.1	192.168.40.1	192.168.20.11
Pattern 1 (30m)	1779 Packets	1242 Packets	1758 Packets	3236 Packets	521 Packets
	7.484	7.124	7.472	8.082	6.256
Pattern 2 (30m)	5942 Packets	1458 Packets	789 Packets	3104 Packets	48 Packets
	8.690	7.285	6.671	8.040	3.871
Pattern 3 (10m)	414 Packets	2691 Packets	1147 Packets	2075 Packets	66 Packets
	6.026	7.898	7.045	7.638	4.190
Pattern 4 (30m)	3544 Packets	3552 Packets	3230 Packets	5780 Packets	654 Packets
	8.173	8.175	8.080	8.662	6.483

TABLE VII
RESULT OF STATISTICAL ANALYSIS.

	Q_1	Q_3	IQR	Threshold	Outlier
Pattern 1 (30m)	7.124	7.484	0.360	6.500	6.256
Pattern 2 (30m)	6.671	8.040	1.369	4.311	3.871
Pattern 3 (10m)	6.026	7.638	1.612	3.247	N/A
Pattern 4 (30m)	8.080	8.175	0.095	7.916	6.483

no actual file-sharing communication has occurred between 192.168.20.11 and 192.168.100.20, and such existence of the file-sharing in a series of communication makes a significant difference to detect as the outlier.

In pattern 2 and pattern 3, illegal communication was conducted only once in the experiment, and no communication was detected as an outlier in pattern 3. In pattern 3, we purposely set the collection period of the packet as a short time. As a possible reason for false negative judgment, therefore, the observation time of pattern 3 was too short to collect enough data for statistical analysis. To verify this reason, when we calculate with the tripled number of legitimate packets assuming the collection period is 30 minutes, the threshold becomes “4.345”, and we can correctly judge the data of the illegal communication “4.190” as the outlier. Besides, the pattern 2, which applied observation time as 30 minutes, although the communication was conducted in similar trends with pattern 3, illegal communication was detected as an outlier correctly.

In pattern 4, though all terminals conducted communication with the same number of times, the difference occurs in the number of packets for the same reason as pattern 1. Hence, this method can judge the necessity of communication correctly.

In summarize, when sufficient data is obtained by observing the packets for a certain period, we can detect illegal communication that is misjudged by the previous CCS by using statistical analysis, and the proposed system judges communication correctly. Thus, the proposed system can make a judgment with higher accuracy than the previous CCS.

However, in these experiments, we did not conduct parameter tuning for deriving threshold, and we can detect outlier correctly as explained in Table VII. Also, the problem of lack of data due to short observation time can be solved by adjusting the parameter. For example, to increase the value of the parameter, we change the parameter from “1.724” to “1.100”. Following this change, each threshold of experiment 2 varies as shown in Table VIII. In this situation, the proposed system detects outlier correctly in all patterns with no false positive.

TABLE VIII
PARAMETER TUNING OF THRESHOLD.

	Threshold (1.724)	Threshold (1.100)	Outlier
Pattern 1 (30m)	6.500	6.728	6.256
Pattern 2 (30m)	4.311	5.165	3.871
Pattern 3 (10m)	3.247	4.253	4.190
Pattern 4 (30m)	7.916	7.976	6.483

In the result of experiment 1, the system displayed a lot of communication judgment between all client terminals, and the router, which is the default gateway of each segment. All these communications are like returned packets. We should not prohibit the returned packets, so these communications should be ignored by the system. We have already pointed out this problem in [1] and listed it as future work to distinguish whether high port communication is legitimate or not.

In the proposed system, the Abnormal Communication Identifier module has all the list of the packet information, and we consider that it can solve the problem. In this paper, we focus on the statistical analysis, and we designed the proposed system in which only communication judged as necessary by the DPort Analysis module is sent to the Abnormal Communication Identifier module to simplify the discussion of the module.

To distinguish the legitimate returned packets and illegal packets, we change the DPort Analysis module to send all results to the Abnormal Communication Identifier module.

When the Abnormal Communication Identifier module receives judgment targets from the DPort Analysis module, it first checks the judgment results. If the judgment result is necessary, it conducts statistical analysis as explained in Section III-C. Otherwise, no statistical analysis should be conducted and put it in the list of judgment targets. Finally, it conducts analysis using the following procedure.

Step 1: Receive all judgment targets that are judged as "Unnecessary" by the DPort Analysis module.

Step 2: Check the destination port number. If it is the well known port or registered port, the target is judged as unnecessary. If it is the dynamic/private port, go to step 3.

Step 3: Check whether there is the outward communication that has the same number in source port as the destination port number of judgment target. If there is no such communication, the target is judged as unnecessary. If there is such communication, go to step 4.

Step 4: Finally, check the found outward communication. If it is judged as necessary, the target is judged as necessary. If it is judged as unnecessary, the target is judged as unnecessary.

VI. Conclusion

In this paper, we extended our previous CCS to solve the problem of it. The previous CCS misjudges abnormal communication as a necessary communication when the destination terminal listens the communication's destination port. We assumed that it is possible to distinguish legitimate communication and abnormal communication by statistical analysis of network traffic volume. Therefore, we adopt a statistical analysis for indicating abnormal communication, which is misjudged as a necessary communication by the previous CCS. We extended the CCS to perform statistical analysis in addition to the previous CCS's analysis. We implemented extended CCS and applied it to a prototype network. In the experiment, the system judged the necessity of communication observed in the network correctly, and we confirmed that the previous CCS problem was solved. As a result, we confirmed the feasibility of the proposed system.

However, because we adopted a statistical analysis, we need a mirrored packet to ensure the significant difference of packet volume between legitimate communication and illegal communication exists. Also, we need to set an appropriate packet collection period according to the traffic volume in the organization's network.

In the experiment, we applied the SDN technique for constructing a network. Our proposal dynamically changes ACL according to the observed traffic, therefore, SDN is one of the best ways to implement the separated network with our method. Constructing an organization network with SDN network equipment is not cost-effective. However, recently SDN has become a more common technology. Recently, there are many kinds of research focusing on SDN technology [28] [29]. Therefore, it can be an expected improvement in the cost of SDN in the near future.

As future works, we will extend the system to apply more complicated environments. Nowadays, the COVID-19 dramatically changes human's work style, and a lot of organizations all over the world adopt working from home. In this situation, many clients connect to the internal network resources from the outside network via VPN and so on. To maintain the security of the organization's network, we have to take into account such outside network devices for constructing a secure internal network.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number JP19K20268. The authors would like to thank Enago (www.enago.jp) for the English language review.

References

[1] Y. Sato, H. Hasegawa, and H. Takakura, "An Evaluation on Feasibility of a Communication Classifying System," *Proceedings of The Thirteenth International Conference on Emerging Security Information, Systems and Technologies*, pp. 9–15, 2019.

[2] P. Cichonski, T. Miller, T. Grance, and K. Scarfone, "Computer Security Incident Handling Guide," *NIST Special Publication 800-61 Revision 2*, 2012, NIST SP800-61 Rev.2.

[3] J. Information-technology Promotion Agency, "Design and Operational Guide to Protect against "Advanced Persistent Threats" Revised 2nd edition," 2011, URL: https://www.ipa.go.jp/files/000017299.pdf [accessed: 2020-12-08].

[4] H. Hasegawa, Y. Yamaguchi, H. Shimada, and H. Takakura, "An Automated ACL Generation System using Directory Service Information and Network Traffic Data (in japanese)," *The IEICE Transactions on Information and Systems (Japanese Edition)*, vol. J100–D, no. 3, pp. 353–364, 2017.

[5] Y. Sato, H. Hasegawa, and H. Takakura, "Construction of Secure Internal Networks with Communication Classifying System," *Proceedings of the 5th International Conference on Information Systems Security and Privacy*, vol. 1, pp. 552–557, 2019.

[6] G. Alessandro, P. Giovanni, C. Alberto, and B. Giuseppe, "Advanced widespread behavioral probes against lateral movements," *International Journal for Information Security Research*, vol. 6, pp. 651–659, 2016.

[7] T. Watanabe, T. Kitazaki, T. Ideguchi, and Y. Murata, "A Proposal of Dinamic VLAN Configuration with Traffic Analyzation and Its Evaluation Using a Computer Simulation (in Japanese)," *IPSJ Journal*, vol. 46, no. 9, pp. 2196–2204, 2005.

[8] A. K. Nayak, A. Reimers, N. Feamster, and R. Clark, "Resonance: Dynamic Access Control for Enterprise Networks," *Proceedings of the 1st ACM SIGCOMM 2009 Workshop on Research on Enterprise Networking*, pp. 11–18, 2009.

[9] T. Miyamoto, T. Tamura, R. Suzuki, H. Hiraoka, H. Matsuo, and et al., "VLAN Management System on Large-scale Network (in Japanese)," *Transactions of Information Processing Society of Japan, IPSJ Journal*, vol. 41, no. 12, pp. 3234–3244, 2000.

[10] N. Gude, T. Koponen, J. Pettit, B. Pfaff, and M. Casado, "Nox: Towards an operating system for networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 3, pp. 105–110, 2008.

[11] "VLAN .Config," 2019, URL: http://www.iiga.jp/solution/config/vlan.html [accessed: 2020-12-08].

[12] M. Mujib and R. F. Sari, "Design of implementation of a zero trust approach to network micro-segmentation," *International Journal of Advanced Science and Technology*, vol. 29, no. 7 Special Issue, pp. 3501–3510, 2020.

[13] N. Wagner, C. Ş. Şahin, M. Winterrose, J. Riordan, J. Pena, D. Hanson, and W. W. Streilein, "Towards automated cyber decision support: A case study on network segmentation for security," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2016, pp. 1–10.

[14] N. Wagner, C. Ş. Şahin, J. Pena, and W. W. Streilein, "Automatic generation of cyber architectures optimized for security, cost, and mission performance: A nature-inspired approach," in *Advances in Nature-Inspired Computing and Applications*. Springer, 2019, pp. 1–25.

[15] M. Akbanov, V. G. Vassilakis, and M. D. Logothetis, "Ransomware detection and mitigation using software-defined networking: The case of wannacry," *Computers & Electrical Engineering*, vol. 76, pp. 111–121, 2019.

[16] A. Bohara, M. A. Noureddine, A. Fawaz, and W. H. Sanders, "An unsupervised multi-detector approach for identifying malicious lateral movement," in *2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS)*. IEEE, 2017, pp. 224–233.

[17] T. Radivilova, L. Kirichenko, D. Ageyev, M. Tawalbeh, and V. Bulakh, "Decrypting ssl/tls traffic for hidden threats detection," in *2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT)*. IEEE, 2018, pp. 143–146.

[18] S. Nakamura, H. Hasegawa, Y. Tateiwa, H. Takakura, Y. Kim, and et al., "A Proposal of Dynamic Access Control with SDN for Practical Network Separation," *IEICE Technical Report*, vol. 117, no. 299, pp. 65–69, 2017.

[19] T. Noro and K. Wada, "A univariate outlier detection manual for tabulating statistical survey (in japanese)," *Research memoir of the statistics*, no. 72, pp. 41–53, 2015.

[20] "MySQL," 2019, URL: https://www.mysql.com [accessed: 2020-12-08].

[21] "Node.js," 2019, URL: https://nodejs.org/ [accessed: 2020-12-08].

[22] "FastAPI," 2019, URL: https://fastapi.tiangolo.com [accessed: 2020-12-08].

[23] "Open vSwitch," 2019, URL: https://www.openvswitch.org [accessed: 2020-12-08].

[24] "Trema," 2019, URL: https://trema.github.io/trema [accessed: 2020-12-08].

[25] "Docker," 2019, URL: https://www.docker.com [accessed: 2020-12-08].

[26] "dpkt," 2019, URL: https://dpkt.readthedocs.io/en/latest/ [accessed: 2020-12-08].

[27] "React," 2019, URL: https://reactjs.org [accessed: 2020-12-08].

[28] F. Kuliesius and V. Dangovas, "Sdn enhanced campus network authentication and access control system," in *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, 2016, pp. 894–899.

[29] F. Nife, Z. Kotulski, and O. Reyad, "New sdn-oriented distributed network security system," *Appl. Math. Inf. Sci*, vol. 12, no. 4, pp. 673–683, 2018.

Proposal of Single Sideband Modulation Scheme with Ideal Low Pass Filter for Wireless Communication Systems

Hiroaki Waraya and Masahiro Muraguchi
Department of Electrical Engineering, Tokyo University of Science
6-3-1 Niijuku, Katsushika-ku, Tokyo, 125-0051, Japan
E-mail: 4319583@ed.tus.ac.jp, murag@ee.kagu.tus.ac.jp

***Abstract* - Recently developed wireless communication systems primarily require a modulation scheme with higher spectral efficiency and higher quality. In this study, Single Sideband (SSB) modulation scheme that transmits two different signals at the same carrier frequency is proposed. Additionally, our proposed method employs a 4-times oversampled digital filter. The distance between the folded spectrum and the carrier frequency is quadrupled, and its filter creates a spectrum with a roll-off rate of zero on the transmission side. A digital filter with such characteristics can help reduce the burden of the subsequent analog filter. Under Additive White Gaussian Noise (AWGN) channel environment, Bit Error Rate (BER) performance of SSB quaternary Amplitude-Shift Keying (4ASK) at roll-off rate of zero is superior by 3 dB with respect to Carrier-to-Noise Ratio (CNR) to 16 Quadrature Amplitude Modulation (16QAM) of the same spectral efficiency, without affecting the effect of extra Hilbert components.**

Keywords - SSB; Hilbert Transform; Multiplexing; Digital Filter; Roll-off Rate.

I. INTRODUCTION

Recently, the demand for wireless communication systems has been increasing with the spread of smartphones, digital terrestrial broadcastings, and wireless Local Area Network (LANs). The frequency resources are being depleted in Ultra High Frequency (UHF) and Super High Frequency (SHF) bands utilized by many wireless systems. So, the high-priority issue for the subsequent wireless systems is a revolutionary modulation scheme with higher spectral efficiency and higher quality. This study proposes SSB modulation scheme with an ideal digital filter to improve the quality of wireless systems.

The SSB system sends data at half the occupied bandwidth when compared to that of the Double Sideband (DSB) system. The SSB signal can be produced by combining the Hilbert transform and quadrature multiplexing, which causes in-phase addition of one sideband and cancellation of the opposite sideband. However, the SSB system is only effective in scalar modulation, such as ASK modulation [1]. In contrast, quadrature modulation, which is a typical DSB modulation, employs two carrier waves of the same frequency that are out of phase with each other by 90°. When the SSB modulation is incorporated with quadrature modulation, the spectrum efficiency is expected to be twice as high as conventional scheme.

Unfortunately, they are not independent of each other as both modulations use the same signal processing for quadrature multiplexing. The in-phase component comprises the I-data and Hilbert transform of Q-data, and the quadrature component comprises Q-data and Hilbert transform of I-data on the receiver side. Thus, lossless demodulation cannot be performed analytically because of extra Hilbert components. In fact, several researchers have recently investigated the SSB Quadrature Phase-Shift Keying (QPSK) modulation scheme. For example, several researches successfully transmitted not only SSB QPSK but also SSB Multivalued Quadrature Amplitude Modulation (M-QAM), using turbo equalization on the receiver side [2]-[5]. However, there are drawbacks that the occupied bandwidth is increased due to the addition of the error correction code, and that their proposed SSB QPSK is sensitive to the residual phase-error under the fading channel because of the effect of extra Hilbert components.

In this study, a single-carrier transmission of SSB 4ASK modulation scheme is proposed. It transmits different signals in Upper Sideband (USB) and Lower Sideband (LSB) without extra Hilbert components, at the same carrier frequency. Under AWGN channel environment, BER performance of the proposed scheme is superior by 3 dB in terms of CNR to the same data rate and the same occupied bandwidth of 16QAM signal. Additionally, the proposed SSB modulation scheme also maintains the good performance in higher order modulation.

In contrast, our proposed scheme employs a 4-times oversampled digital filter to obtain an ideal spectrum. When SSB modulation is applied to a QPSK with a raised cosine filter, a drawback is that the BER performance at low roll-off rate deteriorates even in AWGN channel environment because of extra Hilbert components [6]. However, the proposed digital filter with oversampling produces the ideal Low Pass Filter (LPF) with a roll-off rate of zero, and the BER performance does not change when the digital filter is applied. Additionally, the burden of a subsequent analog filter can be reduced on the transmission side, because the alias can be moved to a position that is four times farther away by oversampling.

The remainder of this paper is organized into as follows: Section II explains about related works. Section III explains how the SSB modulation is performed, and Section IV gives details on digital LPF. Section V presents our proposed SSB

4PAM modulation scheme using digital LPF and describes how two different data on the same carrier frequency are multiplexed. Section VI presents the performance evaluation and simulation results of the proposed scheme. Finally, the concluding remarks of the study are mentioned in Section VII.

II. Related Works

In the field of wireless communication, several studies have conducted to apply SSB modulation, which is a scalar modulation, to a quadrature amplitude modulation such as QPSK. This modulation method is called the Quadrature-Single Sideband (Q-SSB) modulation. The Q-SSB modulation scheme is expected to have twice the spectral efficiency. The time-domain signal of the conventional Q-SSB modulation scheme in the case of USB is expressed as follows:

$$\begin{aligned} S_u(t) = &\left\{I_u(t) + \widehat{Q_u}(t)\right\} cos\, 2\pi f_c t \\ &+ \left\{-\widehat{I_u}(t) + Q_u(t)\right\} sin\, 2\pi f_c t. \end{aligned} \quad (1)$$

On the receiver side, $\left\{I_u(t) + \widehat{Q_u}(t)\right\}$ is extracted in the in-phase component, and $\left\{-\widehat{I_u}(t) + Q_u(t)\right\}$ is extracted in the quadrature component. For example, Fig. 1 depicts the receiving signal $\left\{I_u(t) + \widehat{Q_u}(t)\right\}$ in the time domain. However, the BER performance deteriorates significantly, because extra Hilbert components, i.e., $\widehat{Q_u}(t)$ and $-\widehat{I_u}(t)$, cannot be removed analytically. In a previous research, a turbo equalizer was employed on the receiver side to suppress the degradation of BER performance. In this case, while the received signal with extra Hilbert components is corrected by the equalizer, the bandwidth is increased owing to the addition of the error correction code and it is sensitive to the residual phase-error under the fading channel. On the other hand, $Q_u(t)$ can be extracted via detection of the peak value of Hilbert component as depicted in Fig. 1. However, the BER performance cannot be improved adequately without an error correction code. In this way, when introducing Q-SSB modulation scheme, how to solve the problem of extra Hilbert components is the important issue [2]-[5].

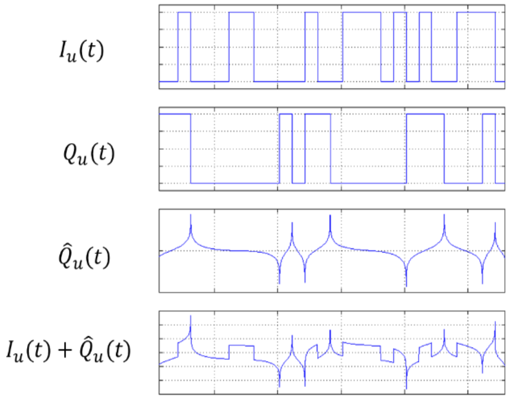


Figure 1. Receiving signal of $\left\{I_u(t) + \widehat{Q_u}(t)\right\}$

On the other hand, when considering a practical Q-SSB modulation scheme with the raised cosine filter at low roll-off rate, the aperture ratio of the eye pattern becomes low due to the presence of extra Hilbert components. Therefore, the BER performance deteriorate as the roll-off rate decreases in not only the fading channel but also AWGN channel, even if turbo equalizer is implemented [6].

As can be seen from the related works described above, the main issues in the Q-SSB modulation scheme are to remove extra Hilbert components and to form a practical spectrum with a roll-off rate of zero without the penalty of BER performance.

III. SSB Modulation

In this section, the performance of the SSB modulation is presented, and the characteristics of SSB modulation and Hilbert transform are explained.

A. Characteristics of SSB modulation

Amplitude modulation (AM) is a technique that multiplies the carrier wave with the information signal and changes the amplitude of the transmission signal in direct proportion to the size of the information signal. The transmission signal $S(t)$ in a general AM method can be represented as follows:

$$S(t) = A[1 + k \cdot m(t)] \cdot cos(2\pi f_c t + \varphi), \quad (2)$$

where, A is the signal amplitude, k is the modulation index $(0 \le k \le 1)$, f_c is the carrier frequency, and φ is the carrier phase. A general envelope waveform of the AM method has an $m(t)$ waveform centered around $\pm A$, the amplitude. If $m(t) = cos\, 2\pi f_m t$, (2) can be rewritten as follows:

$$\begin{aligned} S(t) &= A(1 + cos2\pi f_m t) \cdot cos\, 2\pi f_c t \\ &= A\, cos\, 2\pi f_c t + \frac{A}{2} cos\, 2\pi(f_c - f_m)t \\ &\quad + \frac{A}{2} cos\, 2\pi(f_c + f_m)t. \end{aligned} \quad (3)$$

For simplifying (3), $A = 1, k = 1$, and $\varphi = 0$. In (3), the first term represents the carrier wave component. The second and third terms are components of the information signal $m(t)$. The lower frequency component than the carrier wave, in the second term, constitutes the Lower Sideband (LSB); and the higher frequency component than the carrier wave, in the third term, constitutes the Upper Sideband (USB). Fig. 2 depicts the spectrum of DSB modulation.

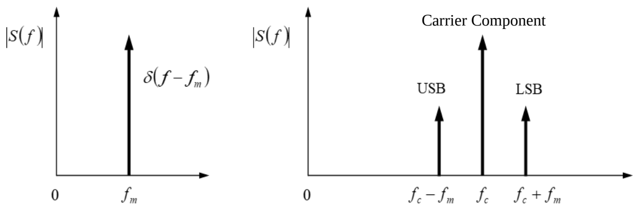


Figure 2. Spectrum of DSB modulation

As depicted in Fig. 2, a spectrum is generated at a location separated by $\pm f_m$ from the carrier frequency f_c. Thus, DSB modulation is a method of transferring an information signal to a carrier band and communicating utilizing the LSB and USB. As observed from (3), the transmission of information is possible either by utilizing the LSB or USB, as they both contain the same information. The transmission of information utilizing only one sideband is called SSB modulation method.

Fig. 3 depicts the SSB transmission spectrum obtained by utilizing the LSB. Here, the negative frequency region is shown as an arithmetic expression, but only the positive frequency region appears as a real signal. In comparison to features of the DSB method, the most prominent feature of the SSB method is that the occupied bandwidth frequency is halved.

B. Hilbert transform

The phase shift method is a way of generating an SSB signal using two 90° phase shifters. The Hilbert transform of a signal $\hat{x}(t)$ is defined as the transform in which the phase angle of all components of the signal is shifted 90° [7]-[8]. It is represented as follows:

$$\hat{x}(t) = H[x(t)] = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x(\tau)}{t-\tau}d\tau = x(t) * \frac{1}{\pi t}. \qquad (4)$$

Here, the signum function is represented as

$$sgn(t) = \begin{Bmatrix} 1, & t > 0 \\ -1, & t < 0 \end{Bmatrix} \Leftrightarrow \frac{1}{j\pi f}. \qquad (5)$$

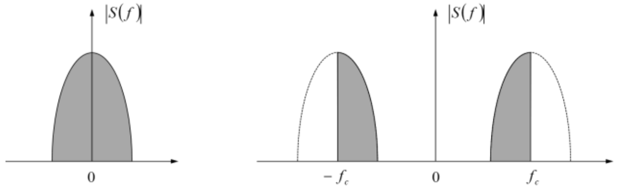
The value of this function in the positive time domain is 1, and in the negative time domain it is -1. On application of the duality theorem to (5), the frequency response of the Hilbert transform $H(\omega)$ is represented as follows:

$$\frac{1}{\pi t} = \hbar(t) \Leftrightarrow H(\omega) = -jsgn(\omega) = \begin{cases} -j, & \omega > 0 \\ j, & \omega < 0. \end{cases} \qquad (6)$$

From (4) and (6), and on application of the convolution in the time domain to the Hilbert transform, the frequency response of the signal transformed by the Hilbert transform is represented as follows:

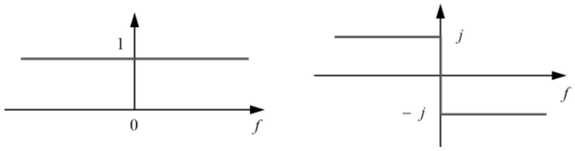
$$\hat{x}(t) = x(t) * \frac{1}{\pi t} \Leftrightarrow \hat{X}(\omega) = X(\omega)\cdot(-j\,sgn(\omega)). \qquad (7)$$

As indicated in (6) and (7), the Hilbert transform delays by 90° at positive frequencies and advances 90° at negative frequencies in the frequency domain. Additionally, the amplitude characteristic is constant regardless of the frequency. Fig. 4 depicts the characteristics of the Hilbert transform.



(a) Base band signal (b) SSB signal

Figure 3. SSB transmission spectrum using LSB



(a) Amplitude characteristic (b) Phase characteristic

Figure 4. Hilbert transform characteristics

The following explains the repeatability of the Hilbert transform. When the Hilbert transform is repeated on the signal that has been processed using the Hilbert transform, the equation is represented as follows:

$$\begin{aligned}\hat{\hat{X}}(\omega) &= \{-j\,sgn(\omega)\}\times\{-j\,sgn(\omega)\cdot X(\omega)\} \\ &= -X(\omega),\end{aligned} \qquad (8)$$

and the inverted original signal is the output. In contrast, if it is a linear transform, the equation is represented as follows:

$$\begin{aligned}H[m(t)\pm n(t)] &= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{m(\tau)\pm n(\tau)}{t-\tau}d\tau \\ &= H[m(t)] \pm H[n(t)].\end{aligned} \qquad (9)$$

Fig. 5 depicts the spectrum transition of the SSB modulation [9]-[10]. In Fig. 5, $S_{USB}(t)\ and\ S_{LSB}(t)$ of the transmitted signal on the frequency domain is represented as follows:

$$S(f) = S_I(f) \pm S_Q(f) = M(f) \pm sgn(f)M(f), \qquad (10)$$

where, $M(f)$ is the modulation signal and $sgn(f)M(f)$ is the Hilbert transform of $M(f)$. In (10) and Fig. 5, when $S_I(f)$ and $S_Q(f)$ are added, SSB modulation by LSB is performed; on the contrary, when they are subtracted, SSB modulation by USB is performed.

On the receiver side, $M(f)$ is extracted at the in-phase component and $\widehat{M}(f)$ is extracted at the quadrature component, as depicted in Fig. 5. As observed from (8), the same signal as the in-phase component can be obtained at the quadrature component upon performing Hilbert transform again. Then, the in-phase and the quadrature components are added to the receiver side.

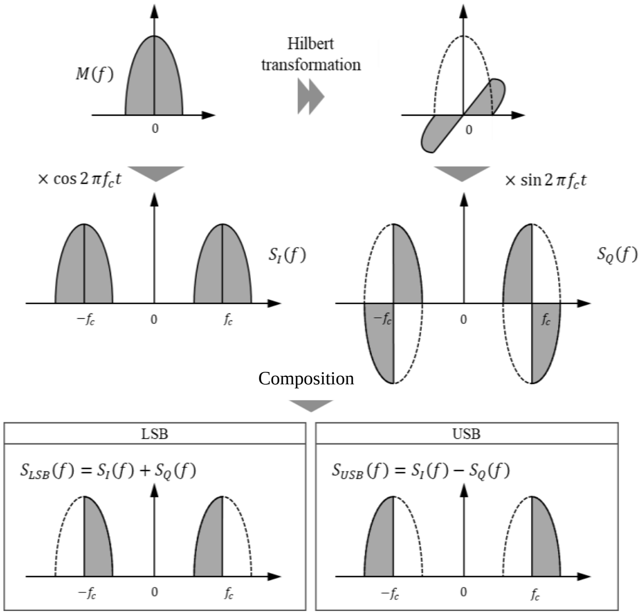


Figure 5. Generating SSB modulation signal

A modulation scheme with such a demodulation scheme is called a Dual Branch SSB (DB-SSB) modulation scheme [11], and it is advantageous in that BER performance is improved by 3 dB. In the DB-SSB demodulation circuit, the variance σ^2 is doubled and the average amplitude E_b of the signal is also doubled in AWGN channel environment by adding the in-phase and quadrature components. Therefore, it can be observed that $|E_b|^2/|\sigma|^2$ of the DB-SSB scheme is superior by 3 dB.

IV. Digital Filter

In this section, the optimal transmission filter design of its modulation scheme is explained.

A. Analog LPF filter

As one of our purposes of this study is to improve the frequency efficiency, the extra occupied bandwidth should be compressed as much as possible. Typically, a spectrum shaping is performed through LPF to prevent the emission of unnecessary radio waves outside the transmission band. The characteristic of a cosine roll-off filter is often considered as a representative of the LPF. The frequency response of the raised cosine filter is represented as follows:

$$H(\omega) = \begin{cases} \sqrt{\frac{1}{2}\left[1 - \sin\left\{\frac{\pi|\omega| - \omega_1}{2\alpha\omega_1}\right\}\right]}, (1-\alpha)\omega_1 \le |\omega| \le (1+\alpha)\omega_1 \\ 1, \qquad |\omega| \le (1-\alpha)\omega_1 \\ 0, \qquad |\omega| \ge (1+\alpha)\omega_1 \end{cases} \quad (11)$$

where, ω_1 is equal to π/T_0. Normally, the Nyquist interval is set identical to the symbol interval. Fig. 6 depicts the frequency response of the raised cosine filter when the roll-off rate is changed. Fig. 7 depicts the impulse response of the raised cosine filter.

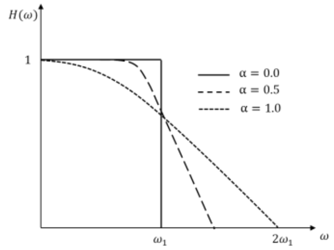


Figure 6. Frequency response of the raised cosine filter

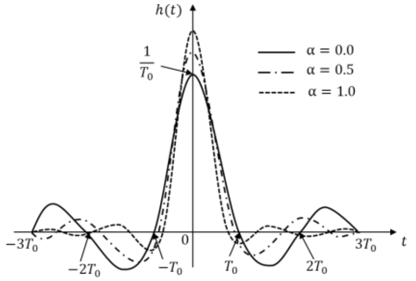


Figure 7. Impulse response of the raised cosine filter

As observed from Fig. 6, the transmission bandwidth becomes wider as α increases, and the bandwidth doubles when α = 1.

The impulse response of the raised cosine filter is represented via an Inverse Fast Fourier Transform (IFFT) being performed on the frequency response. The equation thereof is represented as follows:

$$h(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(\omega)e^{j\omega t}d\omega$$

$$= \frac{1}{1 - 16\frac{\alpha^2 t^2}{T_0^2}}\left[\frac{\sin\left\{(1-\alpha)\frac{\pi t}{T_0}\right\}}{\frac{\pi t}{T_0}} + \frac{4\alpha}{\pi}\cos\left\{(1+\alpha)\frac{\pi t}{T_0}\right\}\right]. \qquad (12)$$

As observed from Fig. 7, the vibration of the impulse response is large, and its response converges slowly, in the case of a steep LPF (α = 0). In contrast, when α is increased, the vibration of the impulse response is small, and its response converges quickly.

If the raised cosine filter, which is a steep LPF (α = 0), is used to improve the frequency efficiency, the side lobe of the impulse response increases and the aperture ratio of the eye pattern decreases. In such cases, the sampling cannot be performed adequately. There is a tradeoff between the amplitude characteristic of the passband and the blocking region of the LPF. Therefore, it is theoretically difficult to completely remove the out-of-band radiation using an analog LPF.

B. Digital Filter with oversampling

As it is difficult to demodulate the receiving signal, in the case of the analog filter of an ideal LPF, the spectrum of the proposed scheme is shaped via digital filter processing. The digital filter with oversampling produces the ideal LPF, with a roll-off rate of zero. Fig. 8 depicts the configuration around the digital filter with oversampling. After the information signal is mapped, a Fast Fourier Transform (FFT) is performed to spread the signal in the frequency domain. The signal on the frequency domain by the FFT is represented as follows:

$$F(k) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} f(t) \cdot e^{-\frac{j2\pi kt}{N}}.\ (k = 0,1,\dots,N-1) \qquad (13)$$

where, $F(k)$ is the frequency-domain signal, and $f(t)$ is the time-domain signal. Then, nulls that are three times FFT length, N, are inserted into the information signal in the frequency domain. As depicted in Fig. 8, the 3N-nulls are inserted, in the middle of the symbol, so that the LSB data and USB data are at both ends of the symbol. Such a processing method is called oversampling. The spectrum thereof is represented in Fig. 9. As observed from Fig. 9, the bandwidth can be compressed to half the original frequency bandwidth. As a result, the spectrum of the signal obtained from the digital filter is in the form of a roll-off rate of zero. Additionally, the alias can be kept sufficiently away from the main lobe, which facilitates subsequent LPF that attenuates the folded spectrum.

The process of inserting 3N-nulls is equivalent to multiplying a rectangular wave in the frequency domain, as depicted in Fig. 9. The signal on the frequency domain after oversampling is represented as follows:

$$F_R(f) = F(f) \cdot rect(f), \qquad (14)$$

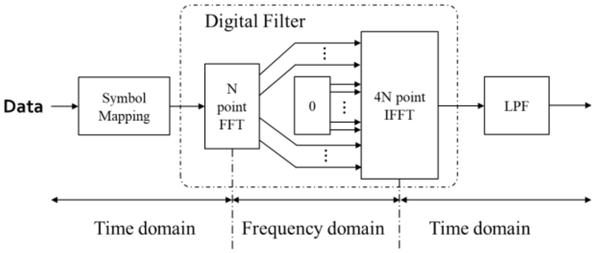


Figure 8. Configuration of the oversampled digital filter

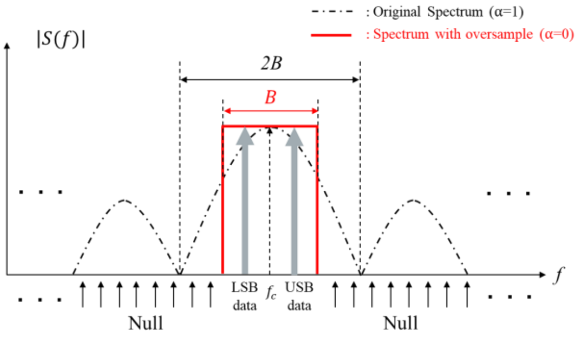


Figure 9. Spectrum with the digital filter

where,

$$rect(f) = \begin{cases} 0, & N < |f| \le 4N \\ 1, & |f| \le N \end{cases}. \qquad (15)$$

The inverse Fourier transform of (15) is represented as

$$\int_{-N}^{N} rect(f) \cdot e^{j2\pi ft} dt = \frac{1}{\pi t} sin\left(\frac{\pi t}{T}\right), \qquad (16)$$

where, N=$1/2T$, and T is the sampling interval. If $f_R(t)$ is the inverse Fourier transform of $F_R(f)$, then

$$\begin{aligned} f_R(t) &= f(t) * \frac{1}{\pi t} sin\left(\frac{\pi t}{T}\right) \\ &= \int_{-\infty}^{\infty} f(t') \frac{1}{\pi(t-t')} sin\left\{\frac{\pi(t-t')}{T}\right\} dt'. \end{aligned} \qquad (17)$$

As $f_R(t)$ is the discrete signal, $t' = nT$. Then,

$$f_R(t) = \sum_{n=-\infty}^{\infty} f(nT) \frac{1}{\pi(t-nT)} sin\left\{\frac{\pi(t-nT)}{T}\right\}. \qquad (18)$$

where $t = kT/4$; IFFT size increases to 4N, which is four-times the oversampled size. As indicated in (18), the time-domain signal after the digital filter processing is represented by the sum of the *sinc* function. Then, the signal that is interpolated between the original signals is generated.

Fig. 10 depicts the time-domain signal after the digital filter processing. The red waveform is the signal before the digital filter processing, and the blue waveform is the signal after the digital filter processing. As depicted in Fig. 10, the oversampling interpolates the part of the transient response between the original signal points, i.e., the quadrupled oversampling interpolates three points between two signal points.

Fig. 11 (a) depicts the eye pattern of an analog steep LPF and Fig. 11(b) depicts the eye pattern, when an LPF is performed after the digital filter processing.

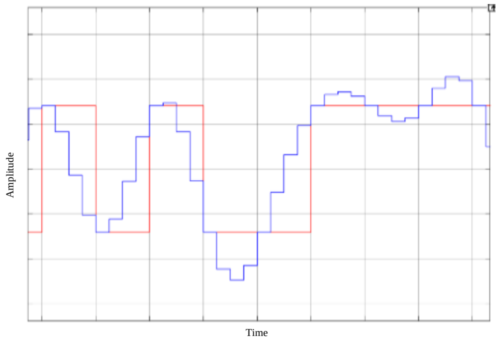
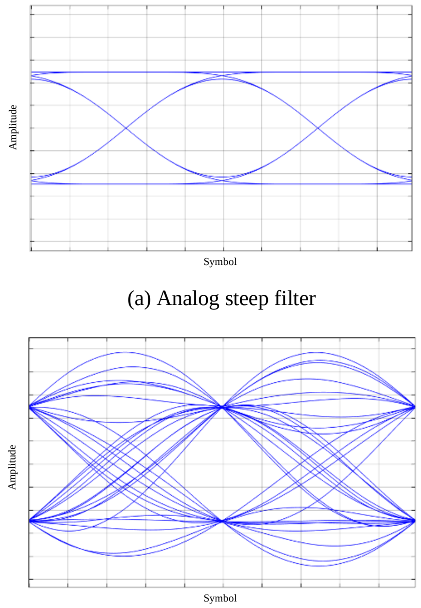


Figure 10. Time domain signal after the digital filter



(a) Analog steep filter

(b) Digital filter

Figure 11. Eye pattern of the signal with each steep filter

It can be seen that the aperture ratio of the eye pattern does not decrease when the digital filter is applied, while the aperture ratio of an analog steep filter, depicted in Fig. 11 (a), decreases.

V. Proposed Method

In this section, the proposed method that uses the SSB modulation is presented, and how two different data of the same frequency are multiplexed is described.

A. *The proposed method*

As a new modulation scheme, a method that adds two data with different amplitudes of the same component is proposed in [1]. The time-domain signal of the SSB method, in the case of USB is expressed as follows:

$$S_u(t) = \left\{I_u(t) + \frac{1}{2}Q_u(t)\right\} cos\, 2\,\pi f_c t \\ + \left\{-\widehat{I_u}(t) - \frac{1}{2}\widehat{Q_u}(t)\right\} sin\, 2\,\pi f_c t. \qquad (19)$$

Fig. 12 depicts the amplitude characteristic of the Q-SSB expressed in (19). Through this method, the I-data and Q-data are extracted on the in-phase component, and Hilbert components of the I-data and Q-data are extracted on the quadrature component. This method is similar to the DB-SSB modulation scheme described in Section III. The BER performance is improved by 3 dB. Regarding the method of separating the I-data and Q-data, the I-data is first obtained by determining whether the value is positive or negative with zero as a threshold value; and the Q-data is obtained by subtracting the I-data from the original signal.

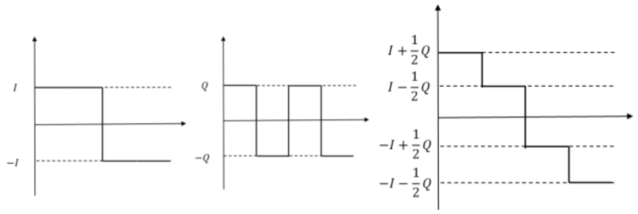


Figure 12. Amplitude of the Q-SSB signal, as in (19)

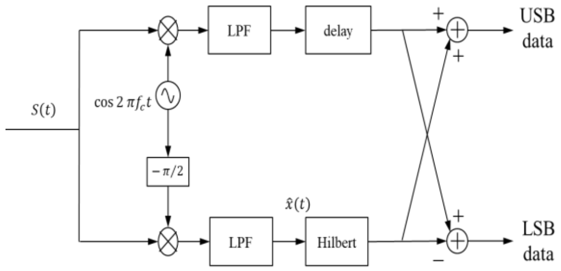


Figure 13. Multiplexed SSB demodulation circuit

However, the signal depicted in Fig. 12 is equivalent to the 4ASK modulation scheme, which increases the amplitude modulation order. Therefore, the 4PAM-based SSB modulation scheme is proposed in this study.

In contrast, our proposed scheme utilizes the digital filter, described in Section IV, which creates a spectrum with a roll-off rate of zero. The burden of the analog filter after the digital filter processing can be reduced on the transmission side.

B. *Multiplexing SSB modulation method*

In the proposed scheme, different information on LSB and USB is transmitted at the same carrier frequency. From (10), the signal generated by multiplexing different information on USB and LSB is represented as follows:

$$S(t) = \{m_u(t) + m_l(t)\}\, cos\, 2\,\pi f_c t \\ + \{-\widehat{m_u}(t) + \widehat{m_l}(t)\}\, sin\, 2\,\pi f_c t, \qquad (20)$$

where, $m_u(t)$ is the signal on USB and $m_l(t)$ is the signal on LSB. Fig. 13 depicts the demodulation circuit for the multiplexed SSB signal. In the upper portion of Fig. 13, $S(t)$ from (20) is multiplied by $cos\, 2\,\pi f_c t$, and passed through the LPF. This signal is represented as in (21). In the lower portion of Fig. 13, $S(t)$ from (20) is multiplied by $sin\, 2\,\pi f_c t$, passed through the LPF, and transformed by the Hilbert transform. That signal is represented as in (22).

$$S_{cos}(t) = m_u(t) + m_l(t) \qquad (21)$$

$$\hat{S}_{sin}(t) = m_u(t) - m_l(t) \qquad (22)$$

Thus, the USB signal is extracted by adding (21) and (22), and the LSB signal is extracted by subtracting (21) from (22).

VI. Performance Evaluation by Simulation

In this study, the spectrum and BER performances of the 4ASK based SSB signal with a digital filter are confirmed. Under the AWGN channel environment, the BER performance of the proposed scheme is superior by 3 dB in the CNR, where compared to the same data rate and the same occupied bandwidth of the 16QAM signal. For accurately analyzing the BER performance of the SSB modulation scheme, the BER performance was measured using E_b/N_0 and CNR. E_b/N_0 is calculated by;

$$E_b/N_0 = CNR + 10log_{10}\left(\frac{T_{sym}}{2T_{samp}}\right) - 10log_{10}(k), \quad (23)$$

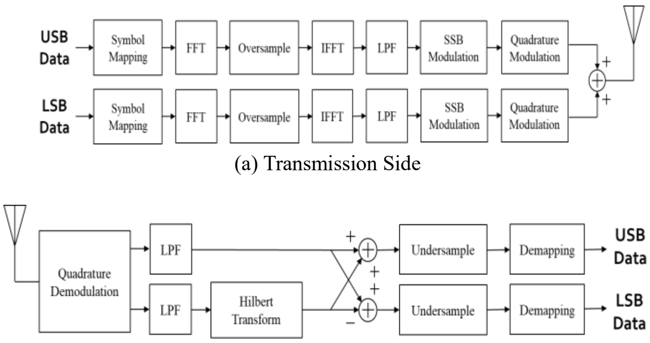
where, T_{sym} is the symbol period, T_{samp} is the sample period, and k is the number of information bits per symbol.

A. Simulation specification

A simulation using MATLAB/Simulink was performed for our proposed method. Table I illustrates the simulation specifications used in this research.

TABLE I. SIMULATION SPECIFICATION.

Parameter	Proposed system	Comparison system
Primary Modulation	4PAM, 8PAM, 16PAM	16QAM, 64QAM, 256QAM
Secondary Modulation	SSB Modulation	Quadrature Modulation
FFT Size	64	64
IFFT Size	256	256
Data Rate	2×2Mbps	4Mbps
Carrier Frequency	40 MHz	40 MHz
Data Size	Single Carrier	Single Carrier
Channel Model	AWGN	AWGN



(a) Transmission Side

(b) Receiver Side

Figure 14. Block diagram of the proposed system

As illustrated in Table I, the 4 Pulse-Amplitude Modulation (4PAM) based SSB modulation is compared with the 16QAM for equivalent bandwidths. Similarly, the PAM-based SSB modulation is compared with the 64QAM and 256QAM to confirm that the advantage of SSB modulation can be utilized even when the proposed scheme is high-ordered. Fig. 14 depicts a block diagram of the proposed system. The undersample on the receiver side extracts data only at appropriate points of the time-domain signal, as depicted in Fig. 10.

B. Simulation results

Fig. 15 depicts the spectrum of the Q-SSB signal that transmits data only at the LSB. Fig. 16 depicts the BER performance in the CNR of the proposed method. Fig. 17 depicts the BER performance in the E_b/N_0 of the proposed method.

As observed from Fig. 15, a part of the opposite sideband is suppressed by the Hilbert transform. It expresses that the Q-SSB signal can be transmitted using only the LSB. As observed from Fig. 16, the SSB method has a BER performance in CNR of 3 dB superior than the DSB 16QAM scheme, which has the same occupied bandwidth. Simultaneously, Fig. 17 depicts that the BER performance of the SSB method, E_b/N_0, is equivalent to that of the DSB 16QAM scheme that has the same spectral efficiency.

Therefore, it is confirmed from Fig. 16 and Fig. 17 that the BER performance of the Q-SSB modulation scheme is superior by 3 dB to that of the 16QAM scheme, depending on the size of the information bit number, k, according to (23). The conventional 4PAM scheme requires only a cosine wave when performing the up-conversion, but the SSB 4PAM scheme also requires a sine wave. Therefore, the signal power of the proposed method on the transmission side is twice as large as that of the conventional 4PAM scheme. This means that the E_b/N_0 of the proposed method deteriorates by 3 dB. However, the proposed method can improve the E_b/N_0 by 3 dB, by implementing the DB-SSB modulation scheme on the receiver side. As mentioned in Section III, the variance σ^2 of AWGN is doubled and the signal average amplitude E_b is also doubled in the demodulation circuit of DB-SSB, by adding in-phase and quadrature components. As a result, the BER performance in $|E_b|^2/|\sigma|^2$ of the proposed scheme is the same as that of the 16QAM scheme with equivalent occupied bandwidth. In terms of BER performance in CNR, the proposed scheme, which is a 2-bit transmission, is 3 dB superior than the 16QAM scheme, which is a 4-bit transmission, from the viewpoint of the number of information bits, k. Such advantages can only be obtained by the discrete signal processing.

Fig. 18 depicts the BER performance in CNR for the 8PAM-based SSB modulation scheme and 16PAM-based SSB modulation scheme. The BER performance of SSB 8PAM is 3 dB superior than that of the DSB 64QAM, which has the same occupied bandwidth. The same is true when comparing 16PAM and 256QAM.

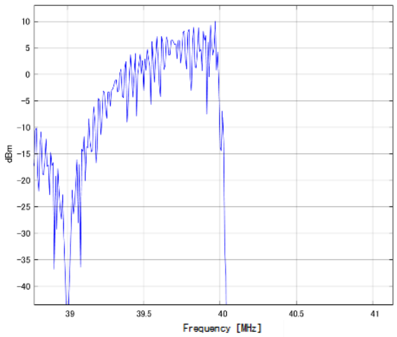


Figure 15. Spectrum of the Q-SSB system

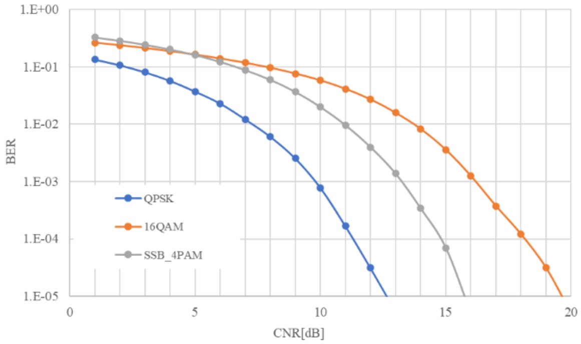


Figure 16. BER performance in the CNR of SSB 4PAM scheme

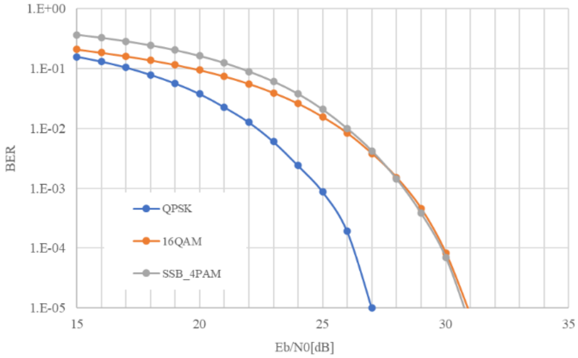


Figure 17. BER performance in the E_b/N_0 of the proposed scheme

From Fig. 16 and Fig. 18, it can be confirmed that the proposed SSB modulation scheme maintains the good performance in higher order modulation, in the AWGN channel environment.

Fig. 19 depicts the spectrum of the multiplexed Q-SSB signal. Fig. 20 compares the BER performance of the Q-SSB signal with data only on the USB and BER performance of the multiplexed SSB modulation scheme. As observed from Fig. 19, each sideband part suppresses the opposite sideband by the Hilbert transform, so that two different 4PAM transmissions can be performed simultaneously on the USB and LSB. In Fig. 20, the multiplexed SSB modulation scheme has the BER performance in E_b/N_0 equivalent to the SSB modulation scheme that transmits data on only USB or LSB.

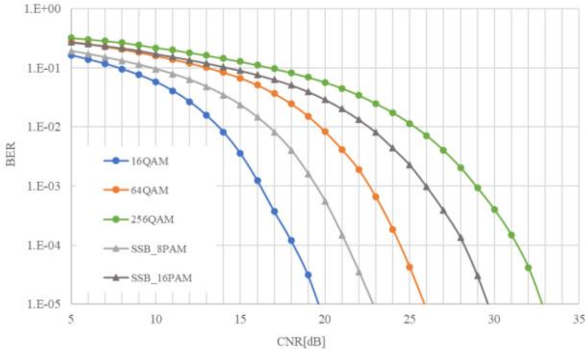


Figure 18. BER performance of SSB 8PAM and SSB 16PAM

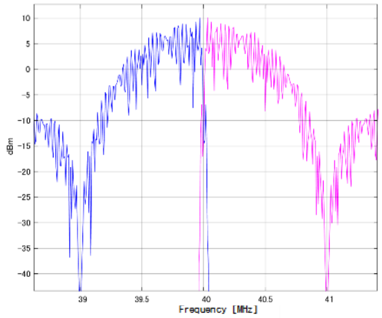


Figure 19. Spectrum of multiplexing SSB modulation scheme

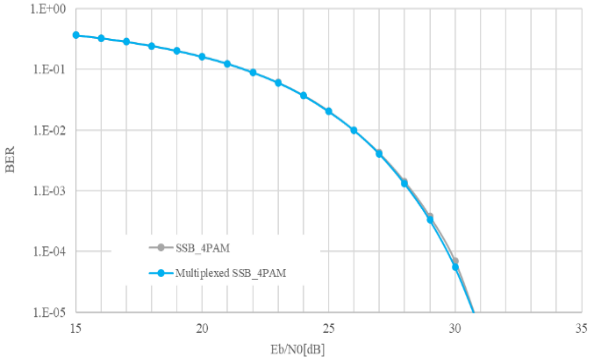


Figure 20. BER performance of multiplexed SSB modulation scheme

The reason why the BER performance does not change even when two signals are multiplexed is that the frequency bandwidth doubles as the number of information bits changes from 2-bit to 4-bit, although the signal power that is added to the signal of the USB and LSB bands doubles on the transmission side at the same time.

Fig. 21 depicts the spectrum of the multiplexed SSB modulation scheme with the digital filter compared to that of the 16QAM modulation scheme. Fig. 22 depicts the folded spectrum of the proposed scheme with a digital filter. Fig. 23 depicts the BER performance of the multiplexed SSB modulation scheme with a digital filter compared to that of the multiplexed SSB modulation scheme without a digital filter.

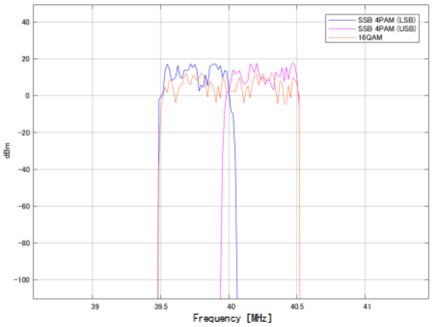


Figure 21. Spectrum of the SSB scheme with the digital filter

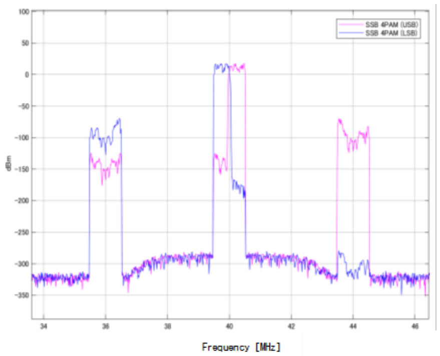


Figure 22. Folded spectrum of the proposed scheme with the digital filter

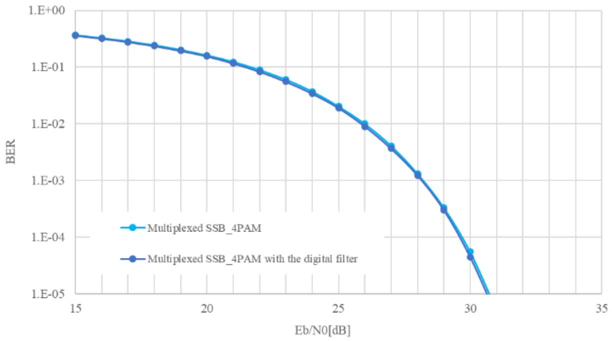


Figure 23. BER performance of the SSB scheme with the digital filter

As observed from Fig. 21, two different SSB 4PAM signals are transmitted adjacent to each other, after a sharp filter processing with a roll-off rate of zero. As a result, the proposed scheme realizes the ideal spectrum on the transmission side with the minimum required bandwidth because the SSB modulation removes the extra band on one side and the digital filter halves the occupied band. The spectral efficiency of the proposed method is equivalent to that of the comparison method, and it is confirmed that it can take advantage of the BER performance without lowering the spectral efficiency.

The frequency of the folded spectrum is 4MHz apart from the carrier frequency of the main lobe when the digital filter is used, as depicted in Fig. 22, and the burden of the LPF that reduces the power of the folded spectrum can be reduced. In Fig. 22, the cutoff frequency of the analog LPF is actually set wide. Without a digital filter, the side lobe adjacent to the main lobe must be reduced by an analog LPF, which requires an LPF with a low cutoff frequency. Thus, there is a high possibility that the data of the main lobe is affected. The cutoff frequency can be widened when a digital filter is applied, so that the demodulation can be easily performed without damaging the signal of the main lobe, as compared to the case with a steep analog LPF having a low cutoff frequency.

In Fig. 23, it is confirmed that the BER performance of the proposed scheme does not change even if a digital filter is applied to the proposed scheme. This is because the proposed scheme extracts appropriate sample points as seen in Fig. 11, on the receiver side. If a steep analog filter is applied, the BER performance deteriorates owing to the signal distortion. Additionally, in the case of the Q-SSB modulation scheme of previous studies as indicated in (1), the BER performance at roll-off rate of zero deteriorates in not only the fading channel but also AWGN channel, owing to the effect of extra Hilbert components [6]. Therefore, it is a great advantage that the BER performance of the proposed scheme with a digital filter that functions as an ideal LPF filter does not deteriorate in AWGN channel environment.

Summarizing the simulation results in Section VI, the multiplexed SSB 4PAM with the digital filter has an ideal spectrum with a roll-off rate of zero, and the BER performance in the CNR is superior by 3 dB compared to the conventional DSB 16QAM, which has equivalent spectral efficiency. The advantage of the proposed scheme is that a spectrum with roll-off rate of zero can be generated without being affected by extra Hilbert components.

VII. Conclusion

In this study, 4ASK-based SSB modulation scheme with a digital filter is proposed. It has been confirmed that under AWGN channel environment the BER performance of the proposed scheme is superior by 3 dB in terms of the CNR to the 16QAM signal for the same data rate and the same occupied bandwidth, and maintains the good performance even in higher order modulation. Additionally, in terms of frequency efficiency, the proposed scheme realizes the ideal spectrum with the minimum required bandwidth without affecting the BER performance. Therefore, in the case of a single-carrier transmission, it is confirmed that the proposed ASK-based SSB modulation scheme is superior in the quality to DSB QAM scheme.

References

[1] H. Waraya and M. Muraguchi, “Proposal of a Quadrature SSB modulation Scheme for Wireless Communication Systems,” The Nineteenth International Conference on Networks (ICN2020), pp. 1-6, 2020.

[2] B. Pitakdumrongkija, H. Suzuki, S. Suyama, and K. Fukawa, “Single Sideband QPSK with Turbo Equalization for Mobile Communications,” 2005 IEEE 61st Vehicular Technology Conference, pp. 538-542, May 2005.

[3] Y. Jiang, Z. Zhou, M. Nanri, G. Ohta, and T. Sato, "Performance Evaluation of Four Orthogonal Single Sideband Elements Modulation Scheme in Multi-Carrier Transmission Systems," 2011 IEEE 74th Vehicular Technology Conference, pp. 1-6, September 2011.

[4] Y. Jiang, Z. Zhou, M. Nanri, G. Ohta, and T. Sato, "Inter-Signal Interference Cancellation Filter for Four-Element Single Sideband Modulation," 2012 75th Vehicular Technology Conference, pp. 1-5, 2012.

[5] A. M. Mustafa, Q. N. Nguyen, T. Sato, and G. Ohta, "Four Single-Sideband M-QAM Modulation using Soft Input Soft Output Equalizer over OFDM," The 28th International Telecommunication Networks and Applications Conference (ITANAC 2018), pp. 1-6, November 2018.

[6] B. Pitakdumrongkija, H. Suzuki, S. Suyama, and K. Fukawa, "Coded Single-Sideband QPSK and Its Turbo Detection for Mobile Communication Systems," IEEE Transactions on Vehicular Technology, VOL. 57, NO. 1, pp. 311-323, January 2008.

[7] X. Wang, M. Hanawa, and K. Nakamura, "Sideband Suppression Characteristics of Optical SSB Generation Filter with Sampled FBG Based 4-taps Optical Hilbert Transformer," 15th APCC, pp. 622-625, 2009.

[8] C. C. Tseng and S. C. Pei, "Design of discrete-time fractional Hilbert transformer," IEEE International Symposium on Circuits and Systems, pp. 525-528, May 2000.

[9] K. Takao, N. Hanzawa, S. Tanji, and K. Nakagawa, "Experimental Demonstration of Optically Phase-Shifted SSB Modulation with Fiber-Based Optical Hilbert Transformers," OFC/NEOEC, 2007.

[10] J. G. R. C. Gomes and A. Petraglia, "A switched-capacitor DSB to SSB converter using a recursive Hilbert transformer with sampling rate reduction," ISCAS 2000, pp. 315-318, 2000.

[11] S. A. Mujtaba, "A Novel Scheme for Transmitting QPSK as a Single-Sideband Signal," IEEE GLOBECOM 1998, pp. 592-597, 1998.

Improving Critical Communications in Northern Canada

Paul Labbé
National Defence
Defence Research and Development Canada
Ottawa, Canada
email: Paul.Labbe@drdc-rddc.gc.ca / Paul.Labbe@forces.gc.ca

***Abstract*—Evolving activities in the Canadian Arctic drive the need for increased dependable high-data rate communication capabilities with low latencies. This study examines performance and challenges of past and current communication technologies available in this northern region for sensor data, video and voice exchange. Some options for operations in the Canadian Arctic are explored accounting for adverse conditions, such as atmospheric disturbances (both natural and man-made) or adversarial attacks on satellites and terrestrial infrastructure. Potential users include Canadian Armed Forces (CAF), North American Aerospace Defense Command (NORAD), off-grid communities and Public Safety, with respective systems requiring machine-to-machine low latency data sharing. Technologies considered include satellites, microwave relays, fiber optic links, radios such as cellular phones, transceivers in high frequency bands (20-30 Mhz), and particularly Unmanned Aerial System (UAS) gateways.**

Keywords-communications; satellite; UAS; Arctic; latency.

I. Introduction

This invited paper expands from [1] to examine telecommunications options for operations in Northern Canada accounting for adverse conditions such as atmospheric disturbances (both natural and man-made) as well for adversarial attacks on satellites and terrestrial infrastructure. While most sections of this paper can be highly scientific, some layman analogies have been added throughout to reach a broader audience.

This Defence Research and Development Canada (DRDC) study was initiated to address Canadian Arctic communication challenges expressed in the new Canada's Defense Policy, *Strong, Secure, Engaged* (SSE) [2], which reaffirmed Canada’s commitment to effective operations in the Arctic. SSE defines an extended Canadian Air Defense Identification Zone (CADIZ), which includes the entire Canadian Arctic Archipelago (Figure 1), in respect of overall North American Aerospace Defense Command (NORAD) modernization efforts towards an improved North Warning System (NWS) requiring high-throughput low-latency communications.

Some aspects of current Northern Canadian communication systems of the Canadian Armed Forces (CAF), off-grid communities, Search And Rescue (SAR), and Public Safety (PS) are identified. A rise in commercial interest, research and tourism in this zone brings increased safety and security demands to address SAR and natural or man-made disasters to which Canada must be ready to respond.

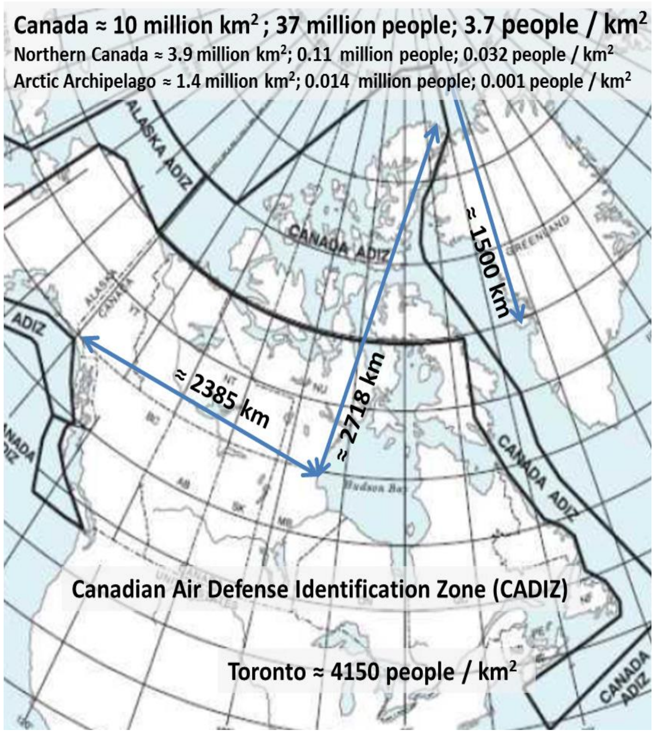

Canada ≈ 10 million km^2 ; 37 million people; 3.7 people / km^2
Northern Canada ≈ 3.9 million km^2; 0.11 million people; 0.032 people / km^2
Arctic Archipelago ≈ 1.4 million km^2; 0.014 million people; 0.001 people / km^2
ALASKA ADIZ
CANADA ADIZ
GREENLAND
≈ 1500 km
≈ 2385 km
≈ 2718 km
ADIZ
CANADA ADIZ
Hudson Bay
Canadian Air Defense Identification Zone (CADIZ)
Toronto ≈ 4150 people / km^2
CANADA ADIZ

Figure 1. North American Canadian operational areas, new Canadian Air Defense Identification Zone (CADIZ), distance vectors and population densities for Canada, Northern Canada, Canadian Arctic Archipelago and Toronto.

Section II of this paper describes the geographic context, climate, topography and size of the area requiring assured communication in all space weather and atmospheric conditions. Section III summarizes relevant aspects of the communication systems envisaged, deployed and used over the last decades in this area, such as Tropospheric Scatter, Geostationary Earth Orbit (GEO) satellite relays at approximately 37000 km above the equator and Medium Earth Orbit (MEO) satellites between 2000 and 36000 km altitudes. Section IV introduces an option using a specific Highly Elliptical Orbit (HEO), the Three APogee (TAP with an apogee around 43000 km altitude over the CADIZ) [3] while Section V brings into perspective novel options offered by several Low Earth Orbit (LEO) satellite constellations at altitudes below 2000 km. Section VI proposes a possible terrestrial architecture with the addition of base stations using towers and/or Unmanned Aerial Systems (UAS) as gateways. Technical considerations begin with Section VII, which compares latencies among the different satellite systems' orbits and other communication technologies. Section VIII examines the pros and cons of some of these options. Section IX explores challenges experienced by radio channels in support of the variety of relevant operations with results from simulations to conclude the scientific analysis. Section X provides a short discussion in layman terms of the presented results and relevant suggestions for further considerations. Section XI concludes with a short summary of findings and recommendations.

II. Context

Northern Canada, aka the North, is the vast northernmost region of Canada variously defined by geography and politics. Politically, the term refers to three Canadian territories: Yukon, Northwest Territories, and Nunavut. This area includes the Arctic Archipelago and covers about 39% of Canada's total land area, and is home to less than 1% of Canada's population.

The Canadian Arctic Archipelago, more succinctly the Arctic Archipelago, groups together all islands lying to the north of the Canadian continental mainland excluding Greenland (an autonomous territory of Denmark). Situated in the northern extremity of North America and covering about 1420000 km^2 (550000 sq mi), this group of 36563 (named) islands in the Arctic Sea comprises much of Northern Canada—most of Nunavut and part of the Northwest Territories. This is about 15% of Canada's geographic area and is home to only about 0.04% of Canada's population. The Arctic Archipelago is experiencing effects of global warming, with some computer models estimating ice melting to contribute a 3.5 cm rise in sea levels by 2100.

CAF's preparedness training and exercises involve several thousand participants and observers, which temporally increase the population density in Northern Canada substantially. For example, Operation Nanook is an annual series of military CAF exercises in the Arctic. It is intended to train different CAF elements (Canadian Army, Royal Canadian Air Force and the Royal Canadian Navy) along other government organizations, such as the Canadian Coast Guard and Royal Canadian Mounted Police in disaster response training and Canadian sovereignty patrols throughout Northern Canada. Another series of exercises, Maple Flag, is conducted south of Northern Canada near CAF Base Cold Lake, which brings about 5000 participants. Both exercises place significant demands on information exchange interoperability, including: voice, data and video, with some telecommunications requiring low latency in order to fulfill machine-to-machine requirements.

A poignant factor in deploying communication systems in Northern Canada is the low population density, which does not support typical commercial business decisions to expand current telecommunications infrastructure the metropolitan Canadian population is accustomed to. With reference to Figure 1, population densities of Europe and India are respectively about 2.3 and 11.6 thousand times greater than those of Northern Canada. It is worth noting that some aspects of evolving 5G technologies are tuned to improve network capacity in high-user density areas like cities. Providing improved services in low density areas will require careful adaptation of current 5G strategies, standards and technologies for deployment in the North.

Summarizing relevant environmental and logistic conditions from [4], the North has extreme weather, with temperatures ranging from -50 to +20 degrees Celsius and wind gusts of up to (hurricane strength) 150 km/h. There is very little precipitation in the Arctic, with an average total of 100 to 200 mm of rain or snow per year. The amount of daylight varies with time of year and with latitude. In Resolute Bay, for example, the sun does not rise above the horizon from early November to early February and does not set from early May to late July.

Permafrost (perpetually frozen ground) is present in most of the Arctic and although typically roughly 10 m thick, it can extend 1 km below the surface. Construction of stable platforms required for large satellite ground stations or microwave towers is costly because the necessary support pillars must be driven below the permafrost; otherwise the platform will shift as the permafrost partially melts during the short Arctic summers (for non-scientific audiences, permafrost shifts like glacier flows). This requires careful site selection to ensure placement where the permafrost is acceptably thin to enable installation preferably into the solid ground underneath. Furthermore, this type of environment is similarly challenging for underground fibre installation.

Year-round access to all communities high in the Arctic is by plane only, increasing the overall cost of travel and accommodation for arctic operations and when planning and installing communications equipment. Communities use seasonal sealifts to transport non-perishable dry cargo (e.g., construction material, household goods, vehicles, etc.) and bulk fuel to them once or twice per year. Large or costly items to transport via air cargo, for example: large aperture satellite dishes or microwave towers, and associated construction materials, must be shipped by sea. This extends the logistic phase of any communication equipment deployment and/or exercise.

III. Practical Arctic Communications Systems

This section summarizes relevant aspects of the following communication systems: Tropospheric Scatter, High Frequency Ground Wave (HFGW), Point-to-Point (P2P) backhaul radio links, Fiber Optic Cable (FOC) and GEO satellite relay systems. Skywave propagation modes are excluded due to their susceptibility to space radiation and atmospheric changes.

Tropospheric Scatter, or Troposcatter, is a beyond-the-horizon communications solution using microwave signal scatter propagating through the troposphere. This phenomenon allows signals to be successfully received around the curvature of the Earth without direct line-of-sight between the transmitter and receiver over vast distances, up to 500 km. Troposcatter systems have been in use for several decades and the technology has seen advances in recent years, including improvements in throughput to 20 Mbps. However, for applications in the North it was found to be too expensive to sustain due to its power demand and infrastructure.

HFGW was documented as an alternate communication network for “nuclear-survivable means of communication for land-mobile missile systems in Europe” [5]. HFGW appears to be a potential alternative means of communication in case of satellite communications disruption, for example due to solar activity or other manmade disturbances. According to International Telecommunication Union (ITU) [6] there are different uses of the terminology and the surface wave is often called the ground wave, or sometimes the Norton ground wave or Norton surface wave, after Norton who developed tractable methods for its mathematical treatment. The generic formulaic expression for the ground wave is the sum of a direct wave, a reflected wave, and a surface wave. When transmitting and receiving antennas are close to the ground, the ground reflection coefficient is -1, and the direct and reflected waves act to cancel each other out, leaving the surface wave as the dominant component. Under such conditions, the ground wave is essentially equal to the surface wave. Empirical results [5] using broadband discone antennas with a cut-off frequency of 19 MHz operating over 20 to 30 MHz near the Arctic Circle between Norway and Germany showed good link connectivity for voice and data communications using narrowband channels for paths over irregular terrain across distances ranging between 19 and 115 km. Based on the empirical results reported, a communication system with its signal spread over 10 MHz with code division multiplexing and sufficient coding gain would offer throughput and fading resilience for high-reliability medium-data-rate channels.

Main challenges in providing microwave backhaul throughout the North using towers include the overall lack of infrastructure, support and staging, the inaccessibility of locations where towers would be built and powering such sites. Microwave links have been extensively used in the North providing reliable connectivity between communities in the Yukon and Northwest Territories that do not have direct access to FOC backhaul. Microwave links provide high throughput with low latency, a significant advantage over GEO satellite systems (below). As microwave frequencies require line-of-sight propagation, towers and topographic features are exploited to extend the range of links beyond the limitations imposed by the curvature of Earth and terrain features along the ground path.

An example of microwave technologies used in remote Arctic locations is the High Arctic Data Communication System (HADCS) of Ellesmere Island, which links Canadian Forces Station Alert (CFS Alert) at latitude 82.5° North (beyond line of site of GEO satellites) to Eureka over a distance of roughly 500 km. The overall communications path includes sending data via a GEO satellite link between Eureka (latitude 79.6° North) and Ottawa 4147 km away. HADCS was retrofitted in 2003 to run entirely on solar power, despite prolonged darkness during winter months. Integrated solar irradiance (W/m^2) over a certain time period and location is called solar radiation, solar exposure, solar insolation, or insolation (J/m^2 or kWh/m^2). Figure 2 shows a HADCS station powered by eight 120 W photo-voltaic (solar) panels arranged in an octagon (eight vertical panels distributed at 45° angular intervals to cover all azimuths).

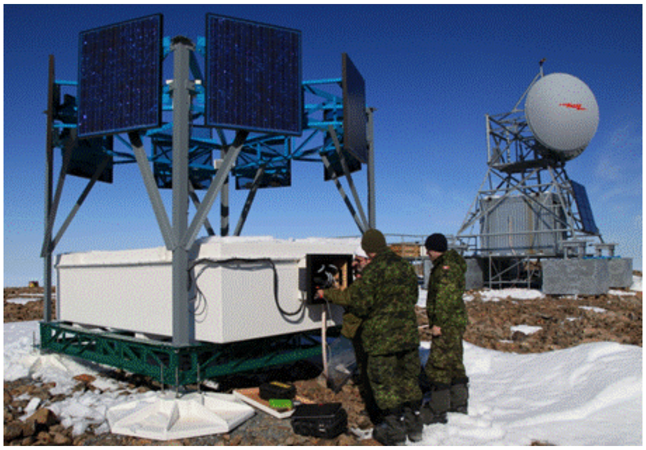


Figure 2. HADCS of Ellesmere Island.

During summer, with 24 hours of daylight, the sun does not dip below the horizon, and all of the solar panels contribute to charging battery banks [4]. Figure 3 shows CFS Alert’s mean daily insolation values expressed in kWh/m^2/day being quite high (Miami, Florida has a value of ~5.26 kWh/m^2/day), making solar a viable option for CFS ALERT during the summer months with sufficient energy stored for the long winter.

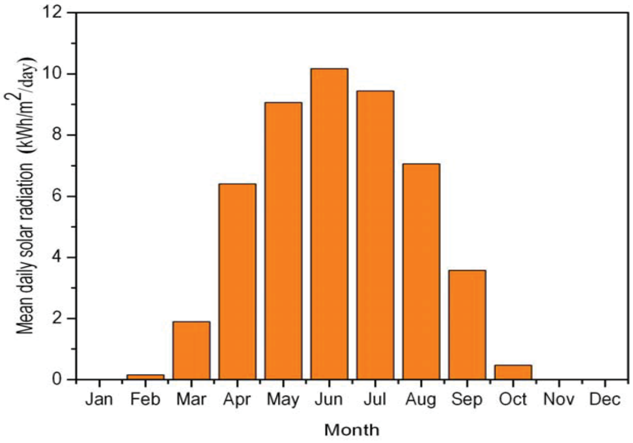


Figure 3. Variability of incident solar power for HADCS.

HADCS comprises seven individual links ranging from 18 to 121 km in length each. The current system operates in bands near 1800 and 2100 MHz, and provides 6.3 Mbps throughput. Helicopter transport is the only means of access to repeater sites [4] as shown in Figure 4.

In the North, costs of construction per kilometer for FOC and microwave links are about equal (approximately $65K/km), when serving a population of roughly 300 people. Above this number of end users, year round, fiber is less expensive [4]. These networks have at least one point of service provisioned via a satellite terrestrial terminal; in practice two or more are used to increase dependability.

Figure 4. A CH-146 Griffon helicopter slings a battery over a repeater site.

GEO satellite systems are well suited for some applications in the North. Given that GEO satellites remain at the same apparent point in the sky when viewed from a corresponding set of locations on Earth having clear line of sight, GEO satellite systems as a whole do not require tracking antennas on the ground as MEO and LEO systems do. This enables the use of lower-cost stationary antennas at ground stations, which is particularly advantageous in harsh environments where moving parts are undesirable. GEO satellite systems provide much of the broadband Internet coverage (including some backhaul connectivity) for communities in the North. Coverage in Nunavut is currently provided by two GEO satellites, Anik F2 and Anik F3, both using the C-band mainly between 4-8 GHz, but also using the 3.7-4 GHz range, which overlaps with the IEEE S-band [4].

However, the use of GEO satellite systems in the North has drawbacks. Due to geometry and the curvature of Earth, there is no clear line-of-sight between equatorial GEO satellites and locations above roughly 80 degrees latitude North as GEO satellites appear under the location’s southern horizon. In practice, communities at latitudes higher than about 70 degrees have elevation-look angles to the GEO satellite below ten degrees as rays emanating from GEO satellite locations are essentially tangential to the ground. This leads to increased absorption and scattering of microwave frequency signals by the atmosphere more generically referred to as absorption, and as signal dissipation at higher frequencies due to precipitation (rain, snow and ice) more generically referred to as rain fade. The five northernmost communities in Nunavut are above 70 degrees latitude and the dominating factor is absorption as precipitation is low overall in the North.

Perhaps the greatest disadvantage of GEO satellite technology use for broadband applications is signal latency: due to the propagation distance to and from the satellite, the minimum delay for a round trip is 480 ms with a median latency of 600 ms. This is an order of magnitude higher than fiber and microwave link latency [4] (details below).

Tests of Mobile Satellite (MSAT) and Iridium capabilities for emergency communications in Northern Quebec showed insufficiencies. The Canadian MSAT system uses GEO satellites, which requires, at such latitude, a high-gain antenna and higher off the ground antenna mounting (resulting in larger installations, location constraints and thus greater costs). These conditions make (emergency) operations difficult and cost-ineffective with very little room to support emergency response.

The US commercial Iridium system with its current handheld, vehicle-mounted and fixed-remote equipment demonstrated different logistic problems, including high cost and poor performance of the handheld telephone sets, which incidentally usually cannot be used inside manmade structures [7].

The Iridium system was designed to be accessed by small handheld telephone sets, the size of a cell phone. Omnidirectional antennas, which are small enough to be mounted on such an Iridium phone, and the low battery power provided, are insufficient to allow the set's radio waves to reach a GEO satellite. In order for such a handheld phone to communicate with them, the Iridium satellites are at LEOs closer to Earth, at about 780 km above the surface. With an orbital period of about 100 minutes an Iridium satellite can only be in view of a handset for about 7 minutes, with the call being automatically "handed off" to another satellite when the previous one passes beyond the local horizon. This requires a large number of Iridium satellites in comparison with GEO satellite based solutions, carefully spaced out in polar orbits, to ensure that at least one satellite is continually in view from every high latitude point on Earth's surface. For seamless coverage at least 66 satellites are required, in 6 polar orbits containing 11 satellites each (with some inactive spares).

For a Canadian Arctic Underwater Sentinel Experimentation (CAUSE) project, transmission of data over a period of 7 months showed that the Iridium Pilot system (with an antenna about six meters above ground) did not fulfill required expectation [8]. “The Pilot data transfer rate for polar transceivers is far less than the advertised rate” [8].

IV. Envisaged HEO/TAP Constellations

Current GEO communication satellites leave the poles uncovered; consequently, Department of National Defence (DND) is exploring options of building the capability or acquiring services from commercial providers with future plans to cover this area. The initial operational capability is tentatively scheduled for 2029.

Quasi-geostationary coverage of Polar Regions can be achieved from HEO satellites. HEO/TAP [3] could be considered under the Enhanced Satellite Communications Project – Polar (ESCP-P) program to provide dedicated, secure and reliable Beyond Line-Of-Sight (BLOS) communications for domestic and continental CAF operations in the Arctic.

In accordance with Kepler's second law, a HEO satellite spends most of the time in the vicinity of apogee (i.e., the farthest point from the Earth's surface, farther than GEO). The orbit could be oriented in such a way that the apogee is over one of the two Polar Regions, so that only two HEO satellites can maintain a continuous view of (presence above) an entire polar zone [9]. When satellite A leaves the service coverage (optimal viewing) zone and heads toward the perigee (i.e., the closest point to the Earth's surface), satellite B rises over the same zone to maintain the same complete circumpolar region in sight. Interestingly, there are periods of several hours per day of coincident (i.e., stereo-like) viewing from the two satellites over most of the circumpolar service area. Such a system could provide meteorological imaging and communication capacity, similar to GEO (including high latency aspects), focused on the polar region. The first HEO satellite system with a period of rotation equal to 12 h called Molniya was implemented for communication purposes in 1965. It established that a two-satellite Molniya HEO constellation can achieve continuous coverage of the polar region 58°–90°N with a Viewing Zenith Angle (VZA) less than 70°. Another HEO system—with a 24-h period—called Tundra is currently used by the satellite Sirius XM Radio service operating in North America. Both orbits, 12-h Molniya and 24-h Tundra, require HEO satellites to be launched with an orbit inclination equal to 63.4°. This value called the critical inclination [9], corresponds to a zero rate of apogee drift due to the second zonal harmonic of the Earth's gravitational field, and ensures a stable position of apogee over the polar zone. If the HEO orbit inclination differs from the critical value, then the apogee gradually drifts toward the equator, requiring orbital maneuvers to maintain the intended orbit position. The farther the orbit inclination is from the critical value, the more resources are required to maintain the orbit. A drawback associated with the 12-h Molniya orbit is the risk linked to hazardous levels of ionizing radiation due to the satellite passing through the Van Allen belts. The highest danger originates from high-energy protons. The Molniya orbit crosses the proton radiation belts at the region of maximum concentration of energetic protons with energies up to several hundred MeV. As an alternative, a 16-h TAP HEO orbit was proposed, providing similar polar coverage as the Molniya HEO system while minimizing the proton ionizing hazards by extending the apogee to 43000 km [3]. The TAP orbit has a ground track with three apogee points repeatable over two days. Such a constellation of two satellites in TAP orbit still revisits ground tracks every 24 h.

V. Perspective of New LEO Constellations

Out of the 11 LEO satellite communications service proposals registered (2014 and 2016) with the US Federal Communications Commission (FCC), the following three are considered based on their maturity [10]: OneWeb, SpaceX (Starlink) and Telesat on Ku (12-18 GHz), Ka (27-40 GHz) and V (40-75 GHz) bands.

To ensure access to affordable high-speed Internet connectivity across rural and Northern Canada, the Government of Canada has invested $85 million and is committed to buy up to $600 million of services over 10 years following launch in 2022 of Telesat's LEO Satellite Constellation [11], which is leveraging Telesat's worldwide rights to ≈4 GHz of Ka-band spectrum.

The analysis in [10] is summarized as follows:

- The maximum total system throughput (sellable capacity) for OneWeb's, Telesat's and SpaceX's constellations are 1.56 Tbps, 2.66 Tbps and 23.7 Tbps respectively.
- A ground segment comprising 42 ground stations will suffice to handle all of Telesat's capacity, whereas OneWeb will require at least 71 ground stations, and SpaceX will need more than 123. And, in this respect, the considerations presented here above regarding ground stations are directly relevant for LEO coverage in the North, particularly satellite tracking antenna functionality.
- In terms of satellite efficiency (ratio between an achieved average data-rate per satellite and its maximum data-rate), Telesat's system performs significantly better (~59% vs. SpaceX's 25% and OneWeb's 22%). This is due to the use of dual active antennas on each satellite, and the lower minimum elevation angle required in their user links.
- OneWeb's system has a lower throughput than Telesat's, even though the number of satellites in the former is significantly larger. The main reasons for this are the lower data-rate per satellite that results from OneWeb's low-complexity satellite design, spectrum utilization strategy, orbital configuration, and payload design, and particularly the absence of Inter Satellite Links (ISLs) to which the analysis presented here will return later.
- If ISLs were to be used in OneWeb's constellation, (even with modest data-rates of 5 Gbps), the number of ground stations required could be reduced by more than half to 27 ground stations.

"To conclude, our analysis revealed different technical strategies among the three proposals. OneWeb's strategy focuses on being first-to-market, minimizing risk and employing a low-complexity space segment, thus delivering lower throughput. In contrast, Telesat's strategy revolves around highly-capable satellites and system flexibility (in diverse areas, such as: deployment, targeted capacity allocation, data routing, etc.), which results in increased design complexity. Finally, SpaceX's system is distinctive in its size; although individually each satellite is not significantly more complex than Telesat's, the massive number of satellites and ground stations increases the risks and complexities of the overall system considerably” [10].

However, although the massive number of satellites and ground stations increases the risks and complexities of the overall system considerably, our experience showed that with appropriate intelligent/adaptive designs, this offers a desirable high level of redundancy and ensures some self-healing capabilities required for dependable critical communication systems supporting vital time-constrained activities and operations, both civilian and military. Overall, employing such massive numbers of satellites might provide room for multifactor synergies beyond the scope of this paper briefly mentioned in the Considerations section X below.

It is worth mentioning that LEO satellite data collected during the 2014-2016 period for the cited 2019 article [10] may have changed considerably from the initial application through the FCC given the fast evolution of regulations and separately of technologies. According to current knowledge, SpaceX made several new applications to evolve the Starlink system design to a multi-constellation with three different orbit layer patterns, in addition to ISLs and a massive number of ground stations to increase connectivity to existing terrestrial infrastructure and creating new infrastructures where none previously existed. However, the current early version of the Starlink satellites do not have ISLs but instead they use advanced antenna arrays [12]. Given SpaceX linkages with National Aeronautics and Space Administration (NASA) and United States Department of Defense (DoD) (e.g., relative to DARPA’s Blackjack program) [13][14], it seems that a possible overarching objective is to deploy capabilities for persistent surveillance, especially to detect and track hypersonic cruise missiles, which provides confidence in considering further synergies in the emergency management and preparedness sphere.

It is worth noticing that Starlink’s beta testing is progressing as reported recently by Canada Satellite [15] More about Starlink progress could be found on the web as per the following references [16][17], as well as reducing interference with space observation with SpaceX's Darksat [17] which exhibited a 50% reduction although much more is required to fulfil the requirement for larger telescopes.

Several developments since the reference publication [10] affect the implementation space. Partly due to COVID-19 and loss of high-risk funding partners, OneWeb went bankrupt in March 2020 while trying to build a satellite constellation to deliver broadband. According BBC news, the UK is part of a consortium with India's Bharti Global, which won a bid to take the company over. Business Secretary Alok Sharma said it would help deliver the "first UK sovereign space capability". The situation is slightly different for Telesat's constellation, which intends to sell part of their licensed radio spectrum in order to generate funds for building their sophisticated LEO satellite technologies with commitments from the Canadian Government [11].

VI. Terrestrial Architectures

UAS gateways (either aerostats, hot air balloon, buoyant gas air balloon, tethered or free-flying, unpowered or powered, dirigibles or high-altitude high-endurance autonomous drones) [4] provide possible communication solutions that merit discussion. The Internet.org consortium has conducted some research into the feasibility of using UAS as communication platforms for remote and underserved locations [15]. Such UAS would be deployed at an altitude of approximately 20 km. By using solar power, UAS systems would be capable of maintaining station above a geographic location, thereby reducing complexity and cost of ground infrastructure when compared to microwave links, without requiring active tracking by the antenna on the ground typically used for MEO and LEO satellite systems. As the UAS would be relatively close to ground, cheaper low-power transmitters could be used, while still enabling high-throughput communications with low latency. UAS would be more susceptible to atmospheric weather than satellites. A previous study [4] assessed that this type of UAS range extenders might have an availability up to 80%. With intervening technology advances, such UAS might be sufficiently reliable today for commercial broadband.

Architecture options offered by new LEO satellite constellations and terrestrial communications, such as UAS, FOC and HFGW given advances in signal processing, multi-beam antennas, spatial diversity and low cost software-defined radios have the potential to substantially improve telecommunication systems availability and reliability in the North. Figure 5 illustrates a hybrid-technology architecture where ISLs and Inter UAS Links (IUASLs) play important roles. Long endurance UAS could use solar with hydrogen fuel-cells. UAS requiring refueling every six months would be ideal (north passage shipping season, tourist season, etc.).

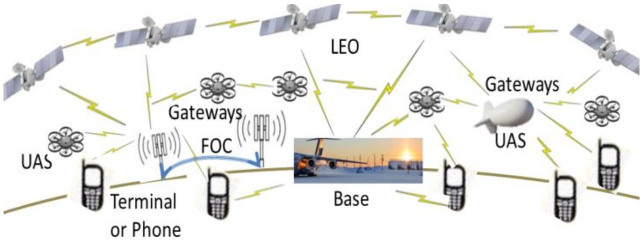


Figure 5. Simplified proposed Northern Communications Architecture.

VII. Latency

Now turning to specific technical aspects, the one way propagation delay of a message is the amount of time it takes to reach its destination from its source. Latency is the delay, usually measured in milliseconds (ms), that occurs in a round-trip data exchange. Round-Trip Times (RTTs) for FOC in large networks [18] using Content Delivery Network (CDN) show useful latency around 18 ms over a distance of 1400 km. HADCS and HFGW offer low latency in this order of magnitude at lower data rates.

High latency typical of GEO satellite links can be very disruptive for some applications, such as video conferencing, and could increase risk in remote health delivery applications including emergency response and particularly remote surgery. Satellite link latencies can also cause low data throughput, caused by the default behavior of communication protocols, which are optimized for shorter distances. A GEO satellite one-way propagation delay is approximately 240 ms due to the large distance between Earth surface location and the satellite; round-trip delivery of a data packet with acknowledgement is approximately 480 ms. This does not include a network delay, which can generally add 50 to 200 ms, depending on where the server is located. GEO satellite systems have a median latency of nearly 600 ms, which includes a median delay of 120 ms incurred by equipment processing speed and network delays in both directions. This makes GEO systems unsuitable to replace cable or fiber systems for applications requiring low latency, particularly impeding machine-to-machine interoperation.

For HEO TAP, with an apogee of 43000 km, the round-trip time is increased by 16% over GEO, i.e., 558 ms or a median latency of 778 ms. Assuming a MEO median orbit of 19000 km, that is 51% of GEO latency, then MEO median latency would be in the order of 306 ms.

The lower orbits of LEO satellites, however, result in latencies much closer to landline quality. The average orbit of the proposed constellations is around 1200 km, an average round trip of 2400 km, incurring a latency of about 31 ms. This is 93.5% improvement over a GEO round trip latency of 480 ms. If the processing speed of LEOs equals that of GEOs then their total median latency would be 151 ms. However, OneWeb tune-up, recorded an average latency of 32 ms in July 2019. As new LEO satellites are designed for high throughput, their overall processing time and network delay must be lower than those used in legacy GEO systems in order to obtain such low latency.

In this paper, since no large sets of empirical results are available for LEO latency, a conservative approach is to use selected simulation results from [19] for an optimal Expected Latency Minimization (ELM) algorithm used in the Software-Defined Networking (SDN) context, which addresses more comprehensively the overall network delay aspect including fading dependence on atmospherics. An interpretation of [19] for its ELM-SDN hypotheses is that LEO's average latency would be around 40 ms with a maximum average latency around 90 ms.

Considering the advantages of LEOs in extending telecommunications coverage and throughput of terrestrial network communications in the North with appropriate gateways, the analysis presented here is extended to include experimental findings.

One challenge of LEO constellations is the frequent handover (also known as handoff) between satellites or between their multiple spotbeams in a satellite footprint. Another challenge stems from terrestrial base station tracking antenna pointing limitations not limited to temperature dependent non-zero slew. These have a good chance to be mitigated for terrestrial base station equipment under low energy and cost regimes by the recent and expected availability of lower-cost components, such as high-speed low-power chip sets and Active Electronically Scanned Array (AESA) integrated boards for microwave systems in Ka and Ku bands. Use of phased array technology is common in satellite radio frequency antennas.

VIII. Pros and Cons of Some of These Options

Next we discuss some radio and LEO channel aspects.

A. *Common radio channel considerations*

Figure 6 illustrates generic terrestrial path loss as function of carrier frequency and shows disadvantages of transmitting at higher frequencies counterbalanced by higher frequencies offering higher throughput. Carrier frequency dependence is also present in satellite communications as mentioned herein.

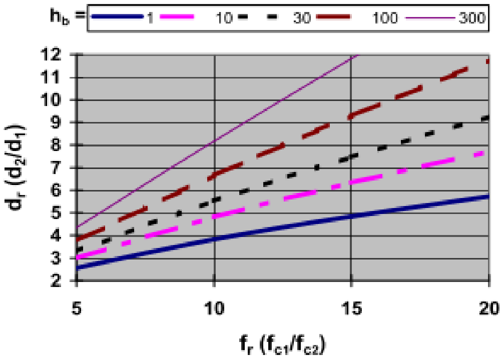


Figure 6. Path loss distance ratio d_2/d_1 as a function of the carrier frequency ratio f_{c1}/f_{c2} for five base station heights and mobile at one meter above ground.

Many of the problems of poor communications performance in remote areas can be alleviated by operating communication networks at lower frequencies. Radio-frequency (RF) passive and active devices, and ancillary components, are generally more efficient when operating frequencies are reduced from 10 GHz to 10 MHz. A similar tendency can be observed in propagation phenomena. Interestingly, urban noise increases as the frequency decreases but this is of less importance presently in the sparsely populated North. Overall, however, effective channel capacity and coverage at 150 MHz is superior to that at 1.5 GHz, at the same time a frequency allocation problem must be addressed (re-allocation of current VHF frequencies is required). The gain in energy-transfer efficiency at lower frequencies can be illustrated by using a propagation prediction model adapted by Hata [20], which estimates the power path loss L_p in dB as:

$$L_p = 69.55 + 26.16\log_{10} f_c - 13.82\log_{10} h_b - Ah_m + (44.9 - 6.55\log_{10} h_b)\log_{10} d \qquad (1)$$

where the carrier frequency f_c is in MHz, antenna heights h are in m and the distance between the antennas *d* is in km. The effective base station height h_b range is from 30 to 300 m: however, in a deployable system for purposes relevant here, mobiles can be used as relays and the lower bound of h_b may be set to one meter. The correction factor $A(h_m)$ in dB is a function of the mobile effective antenna height h_m and size of the city (the Hata model dealt with an urban environment); we neglect this term in our model because the North is sparsely populated, in addition h_m is set to one meter. Applying (1) to two frequencies, f_{c1} and f_{c2} with a ratio $f_r = f_{c1}/f_{c2}$ larger than 1, all the other parameters of (1) being the same, and equating the path loss $L_{p1} = L_{p2}$, d_1 and d_2 become the dependent variables. With $d_r = d_2/d_1$, we obtain the following equation:

$$d_r = f_r^{\left(\frac{26.16}{44.9-6.55og_{10}\, h_b}\right)} \qquad (2)$$

Equation (2) indicates that, for a given path loss, we can increase the distance between the receiver and the transmitter by decreasing the operating frequency of the terrestrial wireless communication network, if other factors are unchanged. In rural areas, foliage and diffraction models for other shadowing effects and surface over-the-horizon radio-propagation allow the derivation of similar equations.

Figure 6 allows estimating the increase in radio coverage when stepping down from 1500 MHz to 150 MHz. Reducing the mobile operating frequency by a factor of 10 extends the communication range by a factor of about 5 for a base station whose effective antenna height is 30 m. If the cell size were 5 km for normal service, it might adaptively extend to 25 km for an emergency temporary service, reducing the logistic burden of covering an area affected by a disaster. Currently, the frequency of 700 MHz is allocated for emergency in Canada.

Colman *et al.* [4] present microwave link systems examples. One is operating at 1.8 GHz and the other at 11 GHz, with similar radiating power. The system at the lower frequency offers a free space maximum range of 333 km while the other, at six times the frequency, qualifies for a free space maximum range of 30 km, which is 11 times shorter. However, the maximum effective throughput rises from 65.4 Mbps to 232 Mbps, which is 3.5 times faster.

Other considerations include the challenge of powering terrestrial systems in the North briefly presented above in Section III. Sources like solar and wind mill power could be combined with sodium-ion batteries, operable at low temperatures.

These considerations apply with due change in particulars to generic mobile units, including airborne, and equally apply to UAS providing cell phone tower functionality.

B. *LEO channel analysis*

Equally, when a satellite is in direct line above a UAS, a ground station or a mobile transceiver, this is the shortest path, the Doppler effect is null, the signal is at its maximum level and usually is less affected by various atmospheric phenomena (cloud, rain, snow fall) aside from possible multiple ground or nearby structures' reflections. The signal is composed of a main direct ray and several secondary reflections. Such a channel displays fades statistics following a Rician distribution.

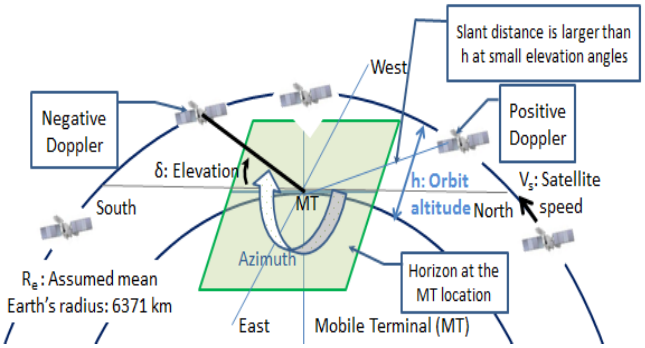


Figure 7. Moving ground or air terminal and LEO satellites.

However, when the satellite ascends (rises) above the horizon (positive Doppler shift) or descends (sets) toward the horizon (negative Doppler shift), i.e., at low elevation angles (δ) above the horizon as shown in Figure 7, the slant range path of the signal is close to the maximum useable distance before being completely shadowed by Earth's curvature. Note that polar orbit LEOs could rise from the North and later from the South over their complete orbit cycle since during a cycle the orbit path turn around Earth (East-West). In the following equations we assume that the mobile terminal is at a negligible height altitude. For UAS and aircraft at high altitudes these equations need to be modified. For a ground terminal, when δ is equal to zero, the slant distance sd can be expressed as follows:

$$sd = \sqrt{h^2 + 2hR_e} \qquad (3)$$

with h = 1200 km and Re = 6378 km, sd = 4100 km. In order to ensure a sufficient signal strength level, LEO satellite constellation designers select a minimum δ large enough to ensure a useful workable Quality of Service (QoS). For example Telesat set δ to a minimum of 20°. Equation (4) [21] provides sd as function of δ:

$$sd = R_e\left(\sqrt{\left(\frac{h+R_e}{R_e}\right)^2 - \cos^2\delta} - \sin\delta\right), \qquad (4)$$

which for δ = 20° gives an sd of 2453 km. Equation (4) can be verified with δ = 90°, which corresponds to the shortest distance value, i.e., sd = h as expected where the Doppler shift is null. Free space Line of Sight (LoS) path loss for each of these distances at 50 GHz (wavelength of 6 mm in the V band) are respectively 198.6 dB, 194.2 dB and 188 dB. Doubling range distance requires a four-fold increase in power for a 6 dB path loss increase at 20° of elevation. It is worst at 0° with about 10 dB or a power ratio of 10 just for the free space loss. However, as the elevation angle δ decreases, more adverse atmospheric phenomena will add to the total path loss, e.g., a cloud may add another 7 dB loss at that frequency.

In addition, the weaker (faint) received signal shows more Rayleigh fading, and is more prone to adjacent channel and jamming signal interference. At small elevation angles, sd is larger, the signal level is lower; the main path competes with relatively strong multipath reflections in addition to being more exposed to atmospheric phenomena over longer distances. Such a channel would likely display deeper and more frequent amplitude fades departing from the Rician distribution, however, following a Rayleigh distribution as hypothesized in [22].

Amplitude measurements of the received radio signal reveal time-varying characteristics resulting from propagation phenomena. When contributions to the total received energy arrive from a large number of reflections, giving a uniform distribution of phases each with similar amplitude, the resulting signal displays Rayleigh amplitude statistics. If a single contribution dominates, the total signal displays Rician amplitude statistics. When contributions arrive from only a limited solid angle around the receiver, the amplitude may follow either a Weibull or a log-normal distribution. In real situations, constant configuration changes as scenarios evolve lead to amplitude statistics that may vary considerably. For example, to address this variety of conditions, the Loo distribution offers an adaptation as function of the elevation angles [23].

The speed of a LEO satellite relative to a fixed ground station V_s can be expressed as function of its orbital period [24] as follows:

$$V_s = \sqrt{\frac{\mu_L}{h+R_e}}. \tag{5}$$

where μ_L is Kepler's constant 398600 km^3/s^2. Consequently the Doppler shift f_{ds} as function of the elevation angle and the carrier frequency [24] is:

$$f_{ds} = \sqrt{\frac{V_s R_e f_c \cos\delta}{C(h+R_e)}}. \tag{6}$$

Using (5) and (6) we find that $f_{ds} = 1059$ kHz for $h = 1000$ km, $V_s = 7.35$ km/s, and $f_c = 50$ GHz. These values represent significant challenges for data links relevant in this present study.

The available time to an initial connection to a LEO satellite, or one of its spotbeams, could be shorter than 3 minutes, hence the frequent need to access a newly visible channel. To address this, in view of challenges with Iridium presented in Section III above, there are a variety of handoff management approaches reported in [25][26]. Some address land mobile systems using terrestrial base stations. Here, we are more concerned with satellite-to-satellite handover [27] and between spotbeams, both inter and intra-satellite spotbeam handovers [28] serving terrestrial fixed or mobile users, base stations, gateways and UAS.

Some definitions from [29] for typical cellular deployments are reused for relevant purposes here. "A hard handover is one in which the channel in the source cell is released and only then the channel in the target cell is engaged. Thus the connection to the source is broken before or 'as' the connection to the target is made—for this reason such handovers are also known as break-before-make. Hard handovers are intended to be instantaneous in order to minimize the disruption to the call. When the mobile is between base stations, then the mobile can switch with any of the base stations, so the base stations bounce the link with the mobile back and forth. A soft handover is one, in which the channel in the source cell is retained and used for a while in parallel with the channel in the target cell. In this case the connection to the target is established before the connection to the source is broken, hence this handover is called make-before-break. The interval, during which the two connections are used in parallel, may be brief or substantial. Soft handovers may involve using connections to more than two cells: connections to three, four or more cells can be maintained by one phone at the same time. The latter is more advantageous, and when such combining is performed both in the downlink (forward link) and the uplink (reverse link) the handover is termed as softer. Softer handovers are possible when the cells involved in the handovers have a single cell site."

Problems with hard handover are the possibility of lost packets, lost requests to repeat them, cost of packet resequencing and consequently additional delays. Soft handover may monopolize more channel resources while being seamless to end users. As reported in [29], there are advanced soft handover approaches that better optimize retention of channel resources and reduce latency.

IX. RADIO CHANNELS TO SUPPORT OPPERATIONS

This section is based on the author's work [30], to which unpublished material is added illustrating the challenges of digital communications between collaborating entities for deployed operations. It illustrates how difficult it is to correctly assess the effective channel capacity in the context of several concurrent operations. The channel capacity is the most stringent factor affecting the performance of distributed information systems made up of mobile-computing nodes communicating via digital radios [31][32][33][34]. In the following analysis it is assumed that the distributed algorithms used are optimal for certain operational loads and observe that optimization is rapidly lost below a certain information exchange threshold [35]. To estimate the performance of such distributed algorithms accurately, several analytical and simulation models supported by experimental results have been proposed in [36][37]. In the majority of cases the models show that spatial and temporal statistical distributions of interrelated phenomena cannot be replaced by mean values without leading to large errors.

In operations such as SAR [38], emergency evacuation and forest-fire fighting, mobile units can be aircraft, helicopters, unmanned vehicles [39], trucks, all-terrain vehicles and backpacks. For this section we assume that information is shared via the radios in order to develop a common operational picture via a distributed database. Each mobile unit automatically reports its position to other participants on pre-established schedules. Decision makers coordinate operations with the aid of computers (accumulated information, geodisplay, decision support) and send appropriate control messages to participants via respective participant communication network nodes. Digital voice competes with the transmission of computer data.

For mobile computing, where purely digital data are exchanged over channels subject to Rayleigh fading, the effective channel capacity (throughput) decreases with an increase in relative velocity between communicating nodes for certain combinations of signal modulation, bit coding and error control protocols. This phenomenon is expected to be exacerbated when applied from terrestrial deployments, where relative velocities are at worst in the 100's km/h, to LEO satellites and UAS gateways where relative velocities are at least an order of magnitude greater.

Next, simulation results are presented illustrating the need to take into account the range of relative velocities encountered between communicating participating mobile computing nodes in order to estimate the performance of the associated distributed systems accurately. Without dismissing other challenges such as tracking the variation of the carrier frequency due to high maneuverability, speed and acceleration of the mobile computing node platform, especially when the Doppler shift changes sign, we focus on the problem of what to do when packet errors occur due to fast-fade phenomena, despite average Signal-to-Noise Ratios (SNRs) being adequate.

We investigate the performance of two different error-control schemes combining certain error-control techniques. The basic scheme uses error detection in a Selective-Repeat (SR) Automatic Repeat-reQuest (ARQ) scheme, SR-ARQ [36]. The other, a hybrid error-control scheme, adds forward-error correction (FEC), SR-ARQ/FEC [40]. For both schemes, Rayleigh-fading channels are assumed for participating nodes moving at relative velocities between 5 and 500 km/h (100 to 1 000 km/h in [41]).

The channel access scheme, a roll-call polling [42][43], is an energy efficient scheme employed in some distributed mobile applications including NATO tactical data exchange LINK-11 and is used here to illustrate aspects of the error-control protocols when sharing fast-aging information. It assumes that a master communication node is trying to gather information from other coordinated participating communication nodes and is the only node controlling channel access. Coordinated nodes wait until they are requested to send their updated data; if they have none, they return only a control packet with an acknowledgment (ACK). If no error-free response is received by the master node after a predefined delay, it polls the next participant node in its list. Reasons for no response include: 1- loss of radio connectivity to the intended participant node, 2- loss of radio connectivity from the participating node to the master, 3- collisions due to a previous loss of radio connectivity, and 4- radio silence either imposed at the queried participating node's platform or due to a fault.

A. *Motivation*

Assuming that many wireless systems will continue to exchange data using existing deployed radios with appropriate modified modes of operation, and/or internal or external upgrades, until new high-performance digital radios become more affordable, it is appropriate to present simulation and analysis results that point out the crucial tradeoffs to be made in link error control, to make best use of the scarce mobile-radio bandwidth. The effective channel capacity and error rate are highly variable and dependent on a variety of environmental factors. Such modified modes of operation, and/or internal or external upgrades, include interfaces to current radios, channel access schemes, error control, signal monitoring, automatic position reporting and interfaces to application computer systems. The selection presented takes into account recent interoperability trends within the telecommunications industry, where two types of circuit-mode services are used for computer-to-computer file transfer [44]:

1. Nontransparent: this service employs a radio link protocol that protects data during the mobile radio-transmission segment (as opposed to transmission over other media such as cable and/or satellite) including ARQ, FEC and flow control. Because of variable conditions in the mobile segment, effective user channel capacity decreases and delay increases as more packets received with errors must be retransmitted to maintain transmitted data integrity. Radio-channel bit-error rate is around 10^{-2} while most applications require data with bit-error rates better than 10^{-6}. This demands adherence to tight requirements on signal modulation, FEC, ARQ and flow-control combinations of such systems.

2. Transparent: this service employs FEC exclusively, i.e., without flow control or ARQ. Users must pre-establish a communication data rate and delay.

Research and development efforts for Asynchronous Transfer Mode (ATM) architecture suitable for mobile computing (civilian or military) may offer an alternative.

Figure 8 shows a model of the signal radiated by a mobile transmitter moving at velocity v_2 and affected by multiple reflectors, scatterers, diffraction layers and other propagation effects (e.g., free-space loss and shadowing), reaching another mobile receiver moving at velocity v_1. When a transmitter moves, its forward wave experiences an increase in carrier frequency and its backward wave a decrease; and at a stationary receiver corresponding positive and negative Doppler frequency shifts are experienced, respectively ($\Delta f_1 = \pm v_1$ / carrier wavelength). Similar effects occur due to a moving receiver: in the direction of the incoming wave front induces a positive Doppler frequency and vice versa. Because of multipath propagation due to scattering, diffraction and reflection, each received signal is represented by a complex vector sum (amplitudes, phases and Doppler shifts). In this paper we assume that the equivalent maximum Doppler frequency to be used in the Rayleigh fading model is the sum of the maximum contributions due to the two mobile platform transceiver velocities v_1 and v_2, that is $\Delta f = \Delta f_1 + \Delta f_2$. By extension, for LEO satellites communicating with a moving air or ground platform, the spreading of the received signal is limited to this value. For this paper we assume that the fade rate depends on the scalar sum of the two mobile (node) transceiver platforms' velocities ($v_1 + v_2$) relative to ground.

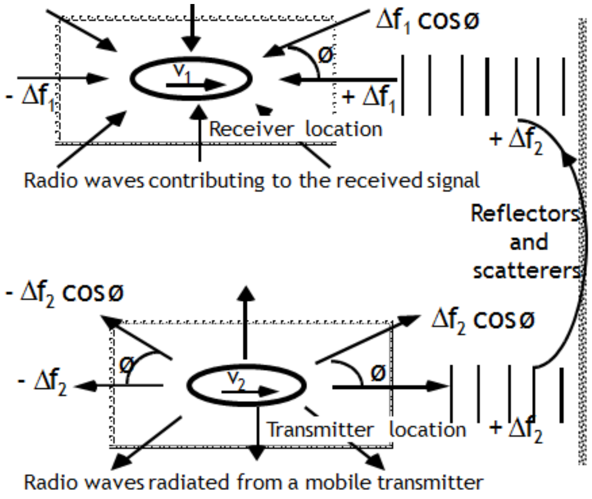

Δf1 cosø
ø
V1
- Δf1
+ Δf1
Receiver location
+ Δf2
Radio waves contributing to the received signal
Reflectors and scatterers
- Δf2 cosø
Δf2 cosø
ø
ø
V2
- Δf2
Transmitter location
+ Δf2
Radio waves radiated from a mobile transmitter

Figure 8. Transmission and reception of wave contributors for moving ground or space platforms.

Free-space path loss, clouds, (rain and/or snow) precipitation shadowing and slow fading, as well as the total noise, determine the local mean value of the observed instantaneous SNR. In the presented model, it is assumed that this SNR local mean is constant over the time required to transmit either: one bit, word, code word or packet. Path losses beyond the Gaussian reference are assumed to be due to fast fading following a Rayleigh distribution.

In [30], we selected a Rayleigh fading model for the fast-fade process, which is not limited to the case, in which adjacent bit errors are independent; the simulation also computes dependent bit errors when the in-fade period overlaps several bits. It is worth noting that the average fade duration is inversely proportional to the combined speeds of the mobile nodes, so it is also related to the Doppler shift. In addition, the number of bits affected depends on the bit rate. In most practical cases this distribution of fades causes higher word-, cell- or packet-error rates than if fades followed a normal, Gaussian or Rician distribution [45][46]. This model also assumes a mobile unit communicating to a base station, moving gateway, including UAS or LEO, or another mobile unit; consequently, the fade rate depends mainly on the mobile unit velocity, or both velocities, and the receiver local mean SNR. In such a case, given constant signal mean and threshold values, the time intervals between fades follow an exponential law and the time duration of fades obeys a Rayleigh law.

B. Error Detection

For most coding schemes we can use the following identification standard: (c, k) NAME-ID, for coding k information bits into c code bits with algorithm NAME-ID. In coding theory, notion of “perfect codes”, which for binary repetition codes, are capable of correcting 0 to (c - 1)/2 errors, i.e., (c, k = 1) codes with only one bit of information since k = 1 (other “perfect codes” include the unique three-error-correcting (23, 12) binary and two-error-correcting (11, 6) ternary Golay codes). In general, these “perfect codes” are not suitable in terms of communication efficiency with unacceptably low code rates (1/c, for binary repetition codes). Error detection in our simulation is assumed to be perfect, i.e., all erroneous streams of c bits are detected. This is not the case for a real-world decoder, whose error-detection capability is limited: for large bit-error rates, the probability of undetected erroneous codes is not negligible. The assumption of perfect error detection suffices for practical purposes here, however, and is current practice in network simulation except when the performance of error-control techniques must itself be assessed specifically.

From simple information theory considerations, for independent error probabilities of adjacent bits, the code-word-error rate [45] without error correction is:

$$P(t=0, e=0) = 1- (1-p_{be})c \qquad (7)$$

In (7) parameter t is the number of errors that a decoder can correct and e is the maximum number of errors acceptable in a received code word of c bits to deliver a correct code word: for decoders without error correction both parameters are zero. The channel bit-error rate is p_{be}. For Binary Phase Shift Keying (BPSK) over Rayleigh fading with 30 dB SNR, p_{be} can be around 0.0003 and for a code of 330 bits, P is 0.006 [46].

In our simulation, perfect error detection capability means that the decoder delivers a correct diagnostic with probability of one for any error pattern even when e = c.

The encoding scheme maps k user-information bits onto code-words of c bits, with c > k. The c-bit streams are designed to be more distinguishable among themselves in noisy conditions than the original k-bit streams, but this encoding increases the necessary transmitted bit rate by c/k.

C. Error Correction

When error probabilities of adjacent bits are independent, the code-word-error rate [45] with an error-correction capability of t code-word bits is:

$$P(t, e \le t) = 1 - (1 - p_{be})^{c} - \sum_{e=1}^{t} \left[\frac{c!}{e!\,(c-e)!} \right] (1 - p_{be})^{c-e} p_{be}^{e} \qquad (8)$$

We maintain the e = c condition for perfect-error detection. However, the condition on e for the decoder to deliver correct-code words is e ≤ t. Because our simulation considers the possibility of non-independent error probabilities of adjacent bits or of any set of bits in a code word, we cannot use (8). However, the conditions e ≤ c for detection and e ≤ t for error correction are valid.

Code-word errors are generated in the simulation according to the Rayleigh statistical fading model using a constant threshold [36]. Given perfect-error detection as indicated above, code words with t and fewer bit errors are corrected. Otherwise for e > t, the word is not corrected by the simulated error-correction decoder; and further actions to correct the word rely on the error-control protocols discussed in the next section. When e > t, real decoders induce new errors with a high probability; and usually the number of errors e' in the delivered word is limited as follows:

$(e - t) \geq e' \geq (e + t)$. We infer from previous work [36] that a word is in error if the sum e of all its bits affected by fading is larger than the decoder error-correction capability: $e > t$.

The highest code rate or code efficiency is 1; this occurs only without coding and overhead. Otherwise, code rates r are smaller than 1 with a number ($b_o > 0$) of overhead bits added to the number k of information bits to give a code-word length $c = k + b_o$. Consequently the code rate is $r = k / c$, and therefore $r < 1$. The bit rate after coding is $r_c = r_s / r = cr_s / k$, where r_s (in bit/s) is the rate, at which k uncoded information bits can be sent. Consequently, the bit rate after coding is larger than before coding for a constant transmission time, $r_c \geq r_s$. This means that if zero error is experienced by a message of k bits, then the message is received correctly at the sink without a need for coding; though that fact is not known in advance. Nevertheless adding coding means that $k + b_o$ bits must be transmitted instead of k; a loss in channel "capacity" proportional to the inverse of the ratio r, i.e., the inverse of the code rate without errors. For clarity, the bit-error rate indicates that only an average of $1 - p_{be}$ of the bits are correct, without the receiver knowing which ones are wrong.

D. Packet

A packet in our model can be defined as one code word (number of code words $n_c = 1$) or multiple code words ($n_c > 1$). The packet length is defined as $L_p = L_d + L_o$, where L_d is the total number of data bits ($L_d = n_c\, c$) and L_o is the total number of overhead bits ($L_o = n_c\, b_o$). More efficient coding can be obtained when error detection and correction are suitably selected, e.g., Viterbi decoding of short-constraint-length convolutional codes with a coding gain of 4.7 dB combined with a high code rate BCH code [46], which is beyond the scope of this paper.

The packet overhead L_o has three overhead components: L_{op} for the protocol, L_{od} for error detection and control, and L_{oc} for error correction alone. In this study, we maintain a sufficiently large L_o to satisfy the assumption of perfect error detection for the two types of packets used: data packets that contain user payload data, and control packets that contain information for node control, channel sensing and packet sequencing. The addition of training sequences may not be necessary since data packets are preceded by one control packet for the roll-call protocol.

According to the above definitions and choice of simulation parameters, we find that the code rate cannot exceed $L_d / (L_d + L_o)$. The code rate is further limited since L_o contains information about the data and some aspect of the previous state of communicating nodes. One advantage of having L_{oc}, L_{od} and L_{op} defined independently in the simulator is the ability to account accurately for overhead even when one of the parameters is set to zero, e.g., when t is zero, there is still some overhead associated with the error detection required by the selective-repeat error-control protocol. This accurate accounting of overhead is essential in measuring the effective channel capacity or user data rate: the amount of user information bits accurately transferred over a period of time in a given set of conditions, e.g., SNR in bandwidth, error-control techniques, modulation, channel-access protocol, environmental and operational parameters, and node load.

E. Protocols used

1) Error-Control Protocols

In the preceding section we discussed two means for controlling errors and information accuracy (code-word-error detection and correction) that are "forward-error-control" techniques, using the forward channel from a source to a sink of information. We further consider the possibility of a feedback channel from the receiver to the transmitter node, permuting these roles in order to feed back cues about transmission success(es) to the source. Error-control protocols are designed to use results of measurements made on immediate and previous observations of information accuracy and signal statistics at destination nodes. These results are then sent back to the source (feedback) where specific outcomes from the perceived accuracy and timeliness of delivered information allow automated counterbalancing actions to be triggered to correct observed situations; actions based on the feedback and based on a specific strategy. Here, error detection involves assessing the accuracy of the received message based on algebraic and probabilistic considerations.

Communicating units—the transmitter and the receiver(s)—store packets and node state variables using unique identifiers in order to distribute common references and to cooperate in the task of transferring user data from node to node correctly. When a packet is rejected by the decoder, an unambiguous request can be made for its retransmission.

For practical reasons (maximum allowable delay, buffer size and sequencing indices) the total number of outstanding packets must be limited. Raising this maximum increases the overhead needed to maintain coherent, unambiguous packet identification among participant nodes in the network. Also, it implies a larger total time to complete a full transmission cycle (the maximum delay in the case of correct delivery of a piece of information after some errors).

For tactical/fast aging data, the end users are generally more concerned with the most recent sensor measurement updates. Since both packet overhead and delivery time must be minimized for concurrent communication efficiency—especially for tactical/fast-aging data [35] over digital radios (less stringent for passive video playback)—as few as 11 bits can be used for sequencing. Other packet overhead bits are required for error detection and correction. Control packets contain a time-history of a receiver's perception of status/information and signal quality and in our study also convey channel-access information.

The resulting maximum value of latency depends on the maximum bit rate and maximum total number of outstanding packets. Practical considerations relevant here to fast aging data include a similarity between some types of tactical data and remote medicine delivery data in the sense that a most recent velocity (computed differences from measured locations) and a most recent heart beat rate (a time average) are important. In each application, the most recent measurement (GPS location / heart beat) can be absent, due to packet loss in transport, if the velocity / heart beat rate are statistically reliable. Receiver's perception can be slightly variable depending on whether such statistics are transmitted from source or computed at receiver. Conversely, in SAR operations all possible available real data is invaluable regardless of radio telecommunications network delay.

2) Selective-Repeat without Forward-Error Correction

ARQ schemes and protocols fall in the feedback-error-control technique family and are by far the oldest and most widely applied error-control protocols in use today [46]. ARQ schemes have three important sub-classes: Stop-and-Wait ARQ (SW-ARQ), continuous *per se* or Go-Back-N ARQ (GBN-ARQ), and continuous ARQ with selected repeats or Selective-Repeat ARQ (SR-ARQ). The last scheme, SR-ARQ, is more efficient with respect to effective channel capacity than the other two [45]. However, it is also the most complex of the three sub-classes.

We selected a basic SR-ARQ strategy with the previously defined perfect-error detector for one set of simulation parameters. The protocol always tries to identify problems with received packets; and a report identifying damaged packets is sent back to the sender. As is the case for any other transmission, this report itself can be disturbed during transmission over the radio channel, the sender may not be informed that a transmitted packet never reached its destination or that the ACK for this packet was disturbed on the return path. We did not limit the simulation to the noiseless feedback channel usually assumed in ARQ performance analyses [46] in order to account for more practical use situations.

In our protocol, we assume that a source tries to send a packet in its queue until the source receives, without error, a confirmation that the packet has been received correctly. At each transmission opportunity, a source node transmits N packets. If N_s of them are not perceived by the source node to be correctly received at the destination node, only N - N_s new packets are taken off the source queue. This decision is made at every transmission opportunity, based on either no confirmation whatsoever, so $N_s = N$, or the correct reception of a control packet indicating N_s packets were determined to be in error by the destination node at its previous reception opportunity.

3) Selective-Repeat with Forward-Error Correction

Combinations of FEC with basic SR-ARQ strategies are called hybrid SR-ARQs. Some improved combinations use a decoding scheme employing multiple copies of a packet that were retransmitted because the first copy was in error. We elected to use FEC on a single packet copy, though this strategy is less efficient than the improved SR-ARQ/FEC, which employs multiple copies of a retransmitted packet (erasures and multiple copies) [37]; in most practical cases the selected scheme is less complex for only a small loss in performance. The important distinction is that in addition to perfect-error detection we add perfect-error correction, for up to t bits of a packet.

As previously, the protocol always tries to identify problems with received packets. If a packet has t or fewer errors, its error status is "no error in packet" and an ACK report is sent towards the source node. This ACK report is piggybacked in a Control Packet (CP in Figure 9) with some FEC, not necessarily a packet of the same length as a data payload packet. In general, control packets can be shorter than data packets. Although shorter packets are less likely to be damaged by an in-fade signal, they remain vulnerable during their transmission over the radio channel. As previously, the sender may not find out that one of its transmitted packets never reached its destination or that the ACK report was disturbed in transport via the feedback channel; with FEC on both types of packets, this happens less frequently than without error correction.

4) Representation of the Simulated Protocols

The protocols simulated are illustrated in Figure 9. We present their process flows by walking through the polling cycle over all the participant nodes (or participating unit labelled PU in the illustration), taking the following steps for each:

1. Master: *select* as current PU the next PU in the network.
2. Master: *read* the ACK from the previous transmission opportunity of this PU and *send* a request based on this ACK with an indication of the number of packets the master will send after this control packet.
3. Master: *send* its packet(s), if any.
4. PU: *send* a control packet including an ACK for the last master packet if it was received correctly, *do nothing* otherwise.
5. PU: *send* packet(s), if there are any and if a control packet was sent in the previous step.
6. Master: *listen* for a reply just after own transmission; if the PU's control packet is received correctly then *record* an ACK for this PU data packet (it may be a zero data packet); otherwise, after waiting one polling period, *record* a **null** ACK for this PU.
7. Master: *continue* cycling (optional end of cycle: *pause* to allow new PUs to join the net), and return to step 1.

F. Selected parameter values for the cases studied

We observed the packet-error rate (our dependent variable) as a function of the combined velocity of the communicating platforms (mobile computing nodes, UAS, LEO satellite), our main independent variable. Although, the simulation can estimate the effect of packet collision for the protocol of Figure 9, we did not select this mode for the results presented for concordance with the material presented in [47]. For similar reasons, we present participating node packet-error rates for a saturated network, that is: when the probability of having N packets in each transmitting queue converges to one or when the normalized offered traffic is equal to or greater than one. For conciseness, we do not present results for user message delay and effective channel capacity as functions of the relative velocity, since they are consistent with the results presented in [47]: they follow the packet-error rate.

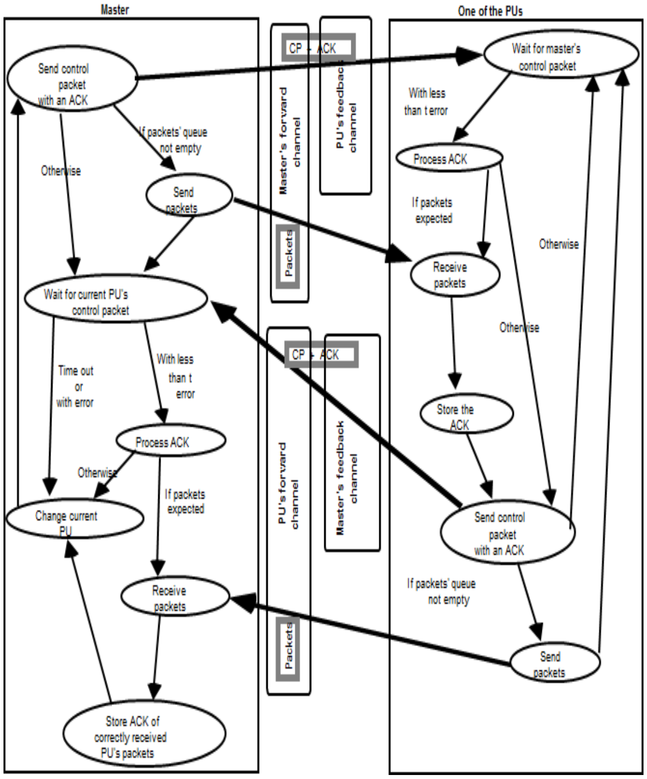


Figure 9. Selective repeat ARQ (with FEC, t ≥ 1) with the roll-call control protocol.

We selected three operational control pairs of values of packet overhead L_o. The first pair simulates a network without error-correction capability, with L_o set to 30 and 20 bits out of 300 information and 200 control bits respectively. In Figure 10, results for this case are labeled t = 0. The two other operational control pairs correspond to networks with FEC capabilities. The simulation uses extraneous bits in addition to the 30 and 20 bits specified previously to increase the overhead by the required number of supplementary bits for correcting t errors: $L_{oc} = 2t$. In Figure 10, results for these cases are labeled t = 2 and t = 3 respectively. The two pairs of L_{oc} for error correction capabilities include one pair with 4 bits for data and 3 bits for control packets, which provide 2 bits and one bit of correction respectively, and another pair of L_{oc} for 6 bits and 4 bits, which provide 3 bits and 2 bits of correction for data and control packets respectively.

Using the selected Rayleigh fading model, high SNRs are associated with small values of the threshold-to-signal ratio ρ. Results for ρ = 0.001 are associated with a SNR of 30 dB in [47].

We assume binary modulation. The symbol period is $T_s = 1 / rc = 50$ μs for a transmission of coded bits at 20 kb/s. We set the values for $v_1 + v_2$ from 36 km/h to 936 km/h. These values are typical of: mobile computing platforms in range of a cellular communications tower or in radio communication range of a UAS. For a wavelength of 0.5 m the carrier frequency is 600 MHz and the maximum Doppler frequency ranges between 20 and 520 Hz. Since the mean value of the in-fade period can be longer than the symbol transmission period, bursts of errors are expected at low Doppler frequencies. The largest packet period is 16.8 ms, which most of the time is shorter than the no-fade interval for ρ = 0.001. For smaller SNR, packets will be affected by multiple fades (multiple bursts of errors); using larger symbols than binary will reduce this effect. The number of information packets N per error-control protocol cycle is set to three for the results presented. However, the model has been investigated by setting N as large as 20. The number of nodes including the master is three for the results presented. The roll-call cycle time has no influence on the results presented and will be used in evaluating the performance of the distributed system.

G. *Information Packet-Error Rate versus the Effective Channel Capacity*

We define the information packet-error rate (P_E) as the ratio of the number of retransmitted information packets relative to the total number of information packets sent including retransmissions. The normalized effective channel capacity or user's data rate (U) can be defined as the ratio of the number of correctly received information bits divided by the total number of bits transmitted:

$$U = \frac{(1 - P_E)(\mathrm{NL_{di}})}{\mathrm{N}\left(\mathrm{L_{di}} + \mathrm{L_{opi}} + \mathrm{L_{odi}} + \mathrm{L_{oci}}\right) + \mathrm{L_{dc}} + \mathrm{L_{opc}} + \mathrm{L_{odc}} + \mathrm{L_{occ}}} \quad (9)$$

The vertical axis on the right of the graph in Figure 10 shows the resulting effective channel capacity for t = 3, which corresponds to 1 - P_E multiplied by 0.73. With t = 0 the multiplier is 0.74. Consequently, there is no need to present separate graphs showing the effective channel capacities for the three multipliers since differences would not be noticeable. The dominant distinctions between the three error-control schemes are accounted for in P_E.

Given a fairly high constant mean SNR (30 dB), Figure 10 confirms the expected dependence of packet-error rates on velocity or Doppler frequency. This commonly observed Rayleigh fading channel condition for mobile radios is the most severe and difficult to overcome (except for Nakagami fading with a parameter smaller than one [48]): it usually causes the highest packet-error rate with a fade rate that increases with the relative velocities of the communicating nodes (mobile units, UAS, LEO satellites, gateways, towers, etc.) Similar results for fast-fading performance [49] support these findings.

In Figure 10, we also provide a typical result for packet-error rate assuming an additive white Gaussian noise (AWGN) channel model for the same SNR. This is a typical result from mean value analyses and valid for many combinations of noise, amplitude variation, modulation and coding. In this type of analysis the fading process is not velocity dependent. The AWGN performance measurements do not display the appropriate distribution statistics for mobile computing.

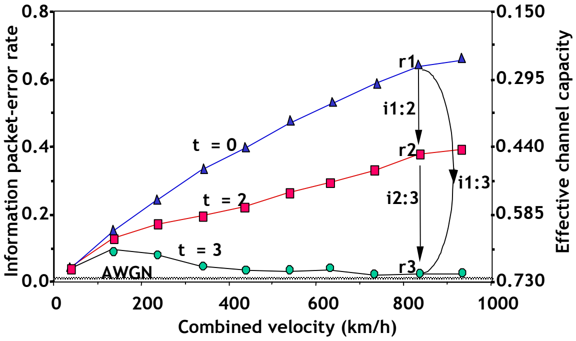


Figure 10. Selective repeat ARQ (with FEC, t ≥ 1) with the roll-call control protocol.

Figure 10 needs to be adapted for higher carrier frequency, bit rate, packet length, code ratio and speed. If all these parameters are proportionally increased, one may observe similar results, i.e., when LEO is at low elevation, the mean SNR will be low (largest slant distance), with the perceived speed and Doppler near its maximum, this corresponds to higher error rates with no FEC (t=0) in Figure 10. While ascending to the shortest distance (equal to the orbit altitude) over a participating node location, a maximum SNR and lower Doppler shift will be attained, as for Figure 10 lowest speed.

H. Power Required for Equivalent FEC Performance

Through simulation we found that an increase in transmit power, labeled **i1:2** in Figure 10, required to bring the packet-error rate indicated by **r1** in Figure 10 to the level indicated by **r2** is **i1:2** = 5 (to decrease the error rate at **r1** to that observed at **r2** we need to increase the current power used by a factor of 5, e.g., from 5 W to 25 W). To obtain the same error rate without FEC, five times the power used with FEC that can correct two errors is required. The power increase, **i2:3**, required to decrease the error rate from **r2** to **r3** is **i2:3** = 332. The power increase, **i1:3**, from **r1** to **r3** is not equal to **i1:2** + **i2:3**; the requirement is **i1:3** = 1250 or from 5 W to 6 kW, a generally unacceptable value for typical long term operation of mobile and handheld units. However, it may be a desirable mode of emergency operation, for example in support of SAR, wherein a powered radio typically operating continuously is switched to a sporadic high power transmission mode to maximize use of an emergency power battery. These three values were estimated in the same way as all other point estimates used to produce the chart in Figure 10 and are consequently based on statistics on random variables. The relation between these variables seems to be **i1:2 • i2:3 = i1:3** with 5 • 332 = 1660 ≈1250. With more samples of this nature we can develop an empirical model expressing the required power ratio as a function of system conditions and parameter values.

Increasing the average SNR by increasing the power of the transmitters to maintain a sufficient data rate is not practical for most mobile applications. A better approach would be to use additional coding and processing gain, e.g., FEC and power spreading. Here we assume that FEC acts to introduce time diversity and spread spectrum as frequency diversity, and both alleviate the impact of high fade rates on packet-error rates as relative velocities increase. For tactical operations, this can provide some level of resistance to jamming through coding to mitigate some level of signal jamming during adversarial attacks. However, with spread spectrum, one can spread the energy of the transmitted signal to a point where it is buried in noise, making it less detectable by basic/legacy electronic support measures systems, although still detectable by more advanced systems. The resulting data rate is higher at high relative velocities but slightly lower at low velocities, making the achievable effective channel capacity more nearly constant over a range of velocities.

The effective channel capacity penalty due to coding is merely a small decrease in the maximum data rate available at negligible relative velocities. Figure 10 shows the total energy saving achievable for the considered system if signal processing, 2 or 3 bit FEC capability, is used instead of higher transmitted power [48]. Using coding and spreading the baseband signal are techniques known to be implementation efficient. For lower SNR and to meet some mobile computing requirements, more coding and processing gain might be required than what is used in our examples.

I. Distribution of Packet-Error Rate Mean Values

We selected the point **r2** from Figure 10 to explore the smoothness of the stochastic processes simulated. As presented in the frequency histogram of Figure 11, there is indication of a good match to a normal distribution of the mean values (point estimates) of the packet-error rate. Note that this normal distribution of mean values of the packet-error rate computed from 600 independent simulation runs seems to be expressed naturally from the statistical processes at play. The temporal distribution of the packets experiencing error(s) cannot be deduced from these results. Accurate performance evaluation of distributed algorithms requires considering distributional dimensions of the underlying systems.

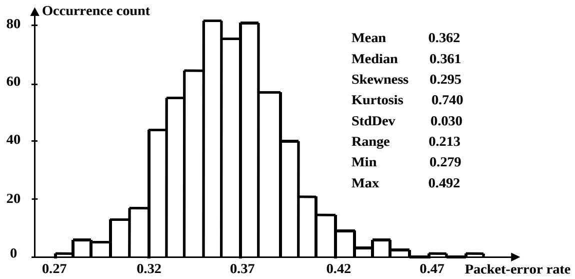


Figure 11. Statistical distribution of point estimates of the packet-error rate found from 600 independent simulation runs.

J. Asymmetry

The improvement for a small FEC capability increase on packets, from 2 to 3 bits, affects error rates experienced by all participant nodes including the master, as indicated by Figure 12. The improvement is larger at higher velocities because of a better match between FEC and error patterns. The in-fade period decreases and the rate of fades increases with velocity though the SNR is constant. The improvement is even larger for the master, since more control packets are involved.

This difference in the packet-error rates between packets sent by the master and by other participant nodes is quite noticeable. Since we selected to not account for collisions, higher master error rates are not due to the larger probability of its packets to collide with PU packets caused by the roll-call protocol asymmetry between a master and participant nodes, but are simply due to the conditional error probability based on the protocol asymmetry. Packets from a participant node are sent to the master only if the appropriate control packet from the master is received correctly, since we ignore bit streams after an incorrectly received master control packet that contains the master reception report concerning the last transmission made by a particular participant node. Consequently, the master reception report on participant node packets is sent on the master-to-participant feedback channel until it finally reaches the concerned participant node without error. The participant node does nothing but wait. For data packets sent by the master the scenario is different. The participant-to-master feedback channel for the reception report concerning master packets does not have this error-free property because each participant node transmits only when it has correctly received an invitation to send data.

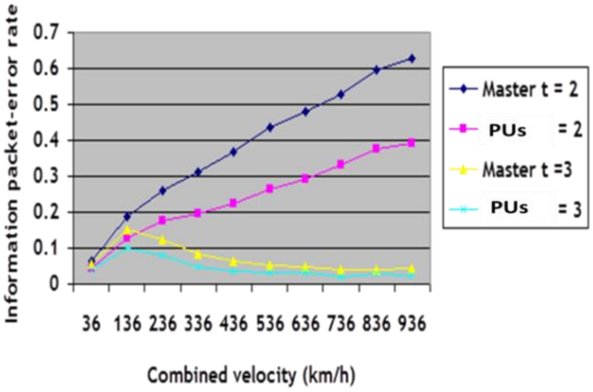


Figure 12. Master's and PUs' information packet-error rates for two values of t.

Despite relatively high average SNR, we observe that the effective rate of correctly transmitted packets drops dramatically as the velocity of communicating nodes increases. The fade rate increases with the velocity while the infade interval is shorter. As the SNR decreases the fade rate increases, but, with longer infade intervals. Both cause an increase in the probability of packet errors on the channel.

K. Summary of the simulation observations

Using mean-value analysis would have precluded noticing the effects of this decrease of effective channel capacity or increase of packet error rate with an increase in relative velocity between all participating nodes. The same is expected for performance analysis of distributed algorithms based on mean-value analysis; parametric empirical models will enable an analyst to visualize the effect of some tradeoffs.

The selected SNR, packet lengths and protocol overhead illustrate that if the design were based only on mean-value analysis, the resulting system would be insufficient to fulfill expected/user operational requirements. Although the effective channel capacity is generally adequate at negligible velocities, the effective channel capacity drops and becomes too small for practical use at relative velocities (ΔV) above 100 km/h. In air-to-air scenarios relative velocities may greatly exceed this range; supersonic speeds often exceed 330 m/s or 1200 km/h.

As the packet error rate increases, the power of the transmitter needs to be increased by a much greater factor to maintain the effective channel capacity and error rate. Typical area-coverage requirements, spectrum management, cost and technology constraints impose on designers the selection of alternatives that combine better modulation, error-control techniques, coding and baseband signal spreading.

In essence, the research results presented here above encapsulate a particular desirable operational parameter space for software defined radios.

X. Considerations

While the results presented here above address technical and scientific aspects of software defined radios, it is hoped that perhaps non-technical audiences find confidence in these results for situation aware development, deployment and coordinated interoperable equipment. In layman terms therefore, the solutions supported by the above analysis amounts to situation adaptable radios.

While costs for tactical operations/scenarios, preparedness exercises and border patrol can be justified in individual jurisdiction (Canada, NORAD), global solutions (NORAD, NATO) can be hard to justify. However, multinational corporations, such as SpaceX and OneWeb (new consortium), without excluding Telesat, are welcome intermediaries particularly in shared fate scenarios:

- Extending the reach of general peaceful operation of internet services to communities in the North remains dominated by commercial factors heavily influenced by population densities and the vast distances between/to Northern communities.
- While not fully sufficient for tactical/border patrol/emergency response/SAR and related preparedness exercises, GEO, MEO and HEO can and do support environmental monitoring and alerting, transactional banking, some provision for remote medicine virtual visits, triage and escalation, while perhaps insufficient for emergency response.
- While typical cellular communications enjoyed in Canadian metropolitan areas may not be available in the North, cellular mobile communications including mobile computing can be employed temporarily in the North in support of preparedness exercises, patrol operations and

prolonged emergencies by employing cellular tower equivalent equipment on airborne platforms including, but not limited to: patrol/SAR recon planes, aerostats (tethered or untethered), dirigibles and particularly on UAS.

- The results presented here particularly show sufficient support for mobile computing in the North via UAS gateways, which when used with the existing terrestrial infrastructure can incorporate, reuse and extend reach. Use of IUASLs can flexibly extend reach further and beyond geographic limitations, at least temporarily (tourist season, shipping season in the Northern Passage as the North Passage becomes more available both due to manmade advances and/or climate change without excluding broader uses) as well in justifying communication infrastructure build-out in the North.
- UAS can be further used based on the results presented here for interoperability as a middle relay layer between LEO satellites and terrestrial infrastructure via software defined radios, which account for high relative velocities between airborne UAS and LEO satellites in outer space. Essentially, by accounting for high Δv, UAS-to-UAS links become substantially equivalent to UAS-to-LEO links for the duration during which a particular LEO satellite is in direct line of sight of a UAS not unlike a cellular mobile set moving within a cell served by a cell tower. Both mobile computing node-to-UAS and UAS-to-LEO radios need to account for corresponding Δv Doppler shifts/fading in similar ways although it is understood that the radios on UAS serving terrestrial mobile communications below are configured differently than the radios enabling space communications above the UAS gateways with LEO satellites.
- Use of phased antenna arrays is promising not only in ground-to-UAS and particularly in direct ground-to-LEO communications to address environmental conditions in the North deleterious to antenna tracking to point to UAS/LEO satellites, but also in simplified UAS designs to handover UAS-to-LEO radio communication channels between UAS space pointing spot beams.
- While low altitude UAS gateway deployments can be susceptible to high winds, the treatise presented here shows balancing gains from reduced precipitation related signal degradation typically plaguing more southern radio communication infrastructure where precipitation is more abundant. Certainly, these results permit selection of UAS operating altitude and therefore retaining interoperability with legacy equipment.

Services provisioned with appropriate parameters and hardware anticipating improvements in coverage, resilience, redundancy, dependability, data rate and low latency include:

- fixed installations like CFS Alert, NWS radar networks, and Forward Operating Bases (FOBs);
- mobiles near fixed installations, airborne or tower gateways or their communication relays;
- short term deployed personnel and platforms for military exercises and operations, emergency operations; and
- off-grid communities of Northern Canada.

Also, there is a need to investigate how Canada could protect space and terrestrial network installation assets.

Satellite transmissions are more susceptible to radiation, jamming and atmospheric disturbances than FOC and over-the-horizon HFGW transmissions. HFGW at 20 to 30 MHz is expected to provide reliable medium throughput for terrestrial communications [5]. FOC offers high throughput and low latency, is commonly deployed around the world and is expanding in Northern Canada [50][51]. However, FOC and HFGW do not offer the area coverage of LEO satellites.

While national direct investment into LEO full constellation deployment remains hard to justify as each of large numbers of LEO satellites spend such a little time over any particular geographic area, the results presented here point to promising and accountable research and development investments.

Further research in mobile computing must be conducted to optimize the sharing of fast-aging data over radio networks (including satellites) used in planning operations critical to our communities, including search and rescue, ice-storm or flood evacuation, etc. Solutions to these problems must be global so that the various organizations involved in such operations—local, national and international—can respond to the needs of the threatened population promptly and cost effectively.

It is self-evident that the move from GEO satellite supported communications to LEO satellite supported communications is a paradigm shift in which:

- Extremely few purpose built, highly redundancy designed, long term operational life expectancy of GEO satellites using long term high cost development cycles are subject to catastrophic failures.
- Extremely large numbers of LEO satellites are built as generic as possible to serve multiple jurisdictions compliant with multiple jurisdictional standards, via software defined radios, for term operation following short term agile development cycles where redundancy is provided by sheer numbers of LEO satellites.

While redundancy has not been explicitly discussed in the technical development in this paper, it bears mentioning that just as the Iridium system makes use of passive pre-deployed Iridium satellite spares, LEO constellations rely on the use of LEO satellite spares. In SpaceX's case, SpaceX's quick low cost turnaround of reusable rocket stages demonstrated just recently, the LEO satellite spares do not have to all be deployed, but a few. The paradigm shift extends from previous hand manufacturing and assembly of GEO satellites to assembly line manufactured LEO satellites and therefore to 'relatively abundant' supply of LEO satellite spares.

Perhaps, beyond the focus of this paper, LEO satellite spares can be very interesting, both the large number of space deployed spares and on the ground LEO satellite spares awaiting launch. Certainly, it can be appreciated that due to low life time expectancy of space deployed LEO satellites, stand-by LEO satellite spares are wasted leaving one to ponder uses for active LEO satellite spares. While certainly, LEO satellite spares are required by the operator to protect investment for commercial ends alone, what is the value of an active LEO satellite spare orbiting over a jurisdiction experiencing a disaster? What is the value of such an active LEO satellite spare for public safety? What is the value of such an active LEO satellite spare in SAR situations regardless of geopolitical borders?

Equally relevant to immediately related fields, what provisions related to the activation of LEO satellite spares can be made in communication license approvals for LEO constellations? Would investment into LEO constellation operations be more palatable, as accountable, for example as insurance premium advance payments equivalent to a rocket launch for 10 to 60 replacement LEO satellites to relieve bandwidth commandeered during extended emergencies? In spite of OneWeb's recent financial troubles, this paradigm shift has made these questions possible.

While the treatise presented here cannot definitely solve all communications shortcomings in the North equally in all situations and in all scenario, UAS gateways are proposed as a middle layer for ground-to-air and air-to-air mobile communication nodes. The obvious question is what is the development and deployment cost and the time horizon of such UAS gateways? The developments presented here show an overlap interoperable not only with legacy equipment while bridging in LEO satellite radio communications in which UAS gateways are no different in radio transmission terms than a LEO satellite constellation layer except for being airborne under. Perhaps another aspect of the paradigm shift in the production of LEO satellites can be appreciated, the assembly line manufacturing can, and often does, produce satellites unfit for space launch due to manufacturing shortcomings of space operational essential components (punctured fuel tanks etc.) making such hardware available for integration into UAS where space essential components are completely unneeded. With this in mind perhaps investment into multinational LEO satellite development and constellation operations can be justified through national acquisition of select such LEO satellite production fractions, at least for research and co-development of UAS gateways.

Aside from tactical considerations supported in this paper, success of LEO satellite enabled communications can more comprehensively address remote medicine delivery in the North including emergency response, perhaps remote surgery. As this paper is being submitted, the implications presented herein are not just thinkable; SpaceX is gathering private interest in LEO internet testing direct to ground.

XI. Conclusion

This article addressed some difficult remaining challenges after many years of communications systems research and development for the North. Findings address DND/CAF challenges to improving capabilities required by future assured communications demands from expected developments in the Arctic and for NORAD operations.

Selected options to improve communications in off-grid areas, more specifically in the North, are expected to provide timely improved shared situational awareness in support of operations where it is currently not well provided, or not available. Such solutions would be revolutionary for our Defense and Security (D&S) capabilities and would progressively provide significant advantages to coalition forces and when CAF operates in collaboration with other Canadian departments including PS and local police in the most demanding emergency and disaster situations.

In the North, if low latency communications are critical for applications or operation objectives, GEO, MEO and HEO satellite systems are insufficient. The least expensive communications systems with low latency in the North should include microwave links, FOC, UAS and LEO systems to ensure fast deployment and access to a large majority of participants in most pressing situations. UAS offers rapid deployment capability on demand in response to PS and CAF situations. LEO/UAS hybrid systems could most definitely extend capabilities of available legacy infrastructure with microwave link and FOC terrestrial infrastructure ensuring connectivity with existing Northern networks and users.

As supported by the results presented here, for moving computing platforms (LEO, UAS, end user vehicle), system designers would have to consider appropriate protocols to match end user requirements for operations. Inevitably, a balance in processing gain, FEC and error control will need to be estimated to match the expected Doppler shift, minimum and maximum relative speed, fast change of in path length affecting free space path loss and atmospheric effects, slow and fast fading, and frequent handovers required.

Overall, the most significant finding is that the advent of low-cost high-performance LEO satellite systems, in conjunction with UAS, can substantially improve communications in the North supporting sustained research and development investment.

Acknowledgment

We are grateful to Marie-Josée Bureau (now at AEG in Montreal) who created the simulation programs critical to this research and provided stimulating and constructive discussions on the approaches and models. We also want to thank Nicholas Bond, one of my former research colleagues, (currently a consultant) for his editorial review, comments and contributions. Finally we thank DRDC for supporting this paper and my wife France Crochetière for her support during this challenging period imposed by COVID-19. Sensitive to current world events, some typical control theory language has been altered herein without loss of essence.

References

[1] P. Labbé, "LEO Satellite Constellations: An Opportunity to Improve Terrestrial Communications in the Canadian Arctic," in *Twelfth International Conference on Advances in Satellite and Space Communications (SPACOMM)*, Lisbon, Portugal, 2020: International Academy, Research, and Industry Association (IARIA), 2020.

[2] Canada: Department of National Defence and Canadian Armed Forces. (2017). *Strong, Secure, Engaged. Canada's Defense Policy*.

[3] A. P. Trishchenko, L. Garand and L. D. Trichtchenko, “Three-apogee 16-h highly elliptical orbit as optimal choice for continuous meteorological imaging of polar regions,” *Journal of Atmospheric and Oceanic Technology,* vol. 28, no. 11, pp. 1407-1422, 2011.

[4] G. Colman *et al.*, “Communications in the North,” Communications Research Centre (CRC), Ottawa, 2014.

[5] J. R. Champion, “An Empirical Investigation of High-Frequency Ground Wave Propagation,” *Johns Hopkins APL Technical Digest,* vol. 13, no. 4, pp. 515-525, 1992.

[6] ITU, “Handbook on Ground Wave Propagation,” in "International Telecommunication Union; Radiocommunication Bureau," R-HDB-59, 2014.

[7] P. Labbé, “GPS and GIS Integration in Mobile Equipment for Improved Mobile Emergency Operations,” in *Proceedings of the 12th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GPS-99); Session A2 - Land Systems & Public Safety*, Nashville Convention Center, Nashville, Tennessee, 1999, pp. 545-555, 1999.

[8] G. J. Heard, “Initial report of CAUSE investigated technologies for remote system deployment,” Defence Research and Development Canada (DRDC) – Atlantic Research Centre, August 2019.

[9] A. P. Trishchenko, L. Garand, L. D. Trichtchenko and L. V. Nikitina, “Multiple-apogee highly elliptical orbits for continuous meteorological imaging of polar regions: challenging the classical 12-h Molniya orbit concept,” *Bulletin of the American Meteorological Society,* vol. 97, no. 1, pp. 19-24, 2016.

[10] I. del Portillo, B. G. Cameron and E. F. Crawley, “A technical comparison of three low earth orbit satellite constellation systems to provide global broadband,” *Acta Astronautica,* vol. 159, pp. 123-135, 2019.

[11] J. Foust. (2020, 11 November) Telesat remains optimistic about prospects for LEO constellation. *SpaceNews*. Available: https://spacenews.com/telesat-remains-optimistic-about-prospects-for-leo-constellation/, Access date: 12 November 2020.

[12] C. Henry. (2019, 15 May) Musk says Starlink “economically viable” with around 1,000 satellites. Available: https://spacenews.com/musk-says-starlink-economically-viable-with-around-1000-satellites/, Access date: 12 November 2020.

[13] S. Erwin. (2020, 11 July) Loft Orbital satellite to carry experiment for DARPA’s Blackjack program. *Spacenews* [News].

[14] S. Erwin. (2019, 26 February) Air Force laying groundwork for future military use of commercial megaconstellations. *Spacenews* [News]. Available: https://spacenews.com/air-force-laying-groundwork-for-future-military-use-of-commercial-megaconstellations/, Access date: 14 Nov 2020.

[15] Canada Satellite, “Starlink Beta Testing Services Terms and Conditions,” ed. Calgary, Alberta, Canada: Canada Satellite, 2020.

[16] M. Sheetz, “SpaceX prices Starlink satellite internet service at $99 per month, according to e-mail,” ed. USA: Consumer News and Business Channel (CNBC), 2020.

[17] J. Tregloan-Reed *et al.*, “First observations and magnitude measurement of SpaceX's Darksat,” *arXiv preprint arXiv:2003.07251*, 2020.

[18] I. N. Bozkurt *et al.*, “Dissecting Latency in the Internet's Fiber Infrastructure,” *arXiv preprint arXiv:1811.10737*, 2018.

[19] W. Cho and J. P. Choi, “Cross layer optimization of wireless control links in the software-defined LEO satellite network,” *IEEE Access,* vol. 7, pp. 113534-113547, 2019.

[20] M. Hata, “Empirical Formula for Propagation Loss in Land-mobile Radio Services,” *IEEE Transactions on Vehicular Technologies,* vol. VT-29, no. 3, pp. 317-325, 1980.

[21] S. Cakaj, B. Kamo, V. Koliçi and O. Shurdi, “The Range and Horizon Plane Simulation for Ground Stations of Low Earth Orbiting (LEO) Satellites,” *IJCNS,* vol. 4, no. 9, pp. 585-589, 2011.

[22] N. Okati, T. Riihonen, D. Korpi, I. Angervuori and R. Wichman, “Downlink Coverage and Rate Analysis of Low Earth Orbit Satellite Constellations Using Stochastic Geometry,” *IEEE transactions on communications* no. tbd, 2020.

[23] S. K. Sharma, S. Chatzinotas and P.-D. Arapoglou, *Satellite communications in the 5G era*. Institution of Engineering and Technology, 2018.

[24] S. Mosunmola B, A. O. Agboola, A. Felix and A. Mohammed, “The Mathematical Model Of Doppler Frequency Shift In Leo At Ku, K And Ka Frequency Bands,” *Published in International Journal of Trend in Research and Development (IJTRD),* vol. 4, no. 5, October 2017.

[25] I. F. Akyildiz, H. Uzunalioğlu and M. D. Bender, “Handover management in low earth orbit (LEO) satellite networks,” *Mobile Networks and Applications,* vol. 4, no. 4, pp. 301-310, 1999.

[26] P. Carter and M. Beach, “Handover aspects for a Low Earth Orbit (LEO) CDMA Land Mobile Satellite (LMS) system,” 1993.

[27] L. Wood, “Internetworking with satellite constellations,” PhD, University of Surrey, 2001.

[28] R. Musumpuka, T. M. Walingo and J. M. Smith, “Performance analysis of correlated handover service in LEO mobile satellite systems,” *IEEE Communications Letters,* vol. 20, no. 11, pp. 2213-2216, 2016.

[29] S. Das *et al.*, “Location Manager based Handover Method for LEO Satellite Networks,” *International Journal of Computer Applications,* vol. 44, no. 12, pp. 43-49, 2012.

[30] P. Labbé, “Mobile Networks: the Effect of Relative Node Velocity on Effective Channel Capacity,” in *(IEEE) 1999 International Conference on Telecommunications, ICT ‘99; Resource Management*, Cheju, Korea, 1999, pp. 147-156: IEEE, 1999.

[31] D. Chirca, “STANDARD TACTICAL DATA LINKS,” *Land Forces Academy Review,* vol. 15, no. 3, p. 393, 2010.

[32] M. Dinc, “Design considerations for military data link architecture in enabling integration of intelligent unmanned air vehicles (UAVs) with navy units,” in *NATO SCI Conference, Germany*, 2009, 2009.

[33] V. Jodalen, R. Otnes and B. Solberg, “Nato standards for HF communications: an overview and technical description,” 2004.

[34] M. Jovanović and R. Todorović, “Joint tactical radio–aspects of standardization,” *Scientific Technical Review,* vol. 53, no. 2, pp. 72-79, 2003.

[35] P. Labbé and R. Proulx, “Model-Based Measures for the Assessment of Engagement Opportunities Implementation and Test Results,” Defence Research Establishment Valcartier (Québec), DREV - R - 9807, November 1998.

[36] R. Comroe and D. Costello, “ARQ schemes for data transmission in mobile radio systems,” *IEEE Journal on*

Selected Areas in Communications, vol. 2, no. 4, pp. 472-481, 1984.

[37] P. Yu, X. Wang and H. Yu, “A novel ARQ scheme applied to wireless communication,” in *2014 IEEE 5th International Conference on Software Engineering and Service Science*, 2014, pp. 1044-1047: IEEE, 2014.

[38] R. E. Hundt and L. Irving, “Final report of the public safety wireless advisory committee,” *Report, PSWAC, September,* 1996.

[39] G. Bevacqua, J. Cacace, A. Finzi and V. Lippiello, “Mixed-initiative planning and execution for multiple drones in search and rescue missions,” in *Twenty-Fifth International Conference on Automated Planning and Scheduling*, 2015, 2015.

[40] E. Weldon, “An improved selective-repeat ARQ strategy,” *IEEE Transactions on Communications,* vol. 30, no. 3, pp. 480-486, 1982.

[41] E. Biglieri, M. Sciuva and V. Zingarelli, “Modulation and coding for mobile radio communications: channels with correlated Rice fading and Doppler frequency shift,” *IEEE transactions on vehicular technology,* vol. 47, no. 1, pp. 133-141, 1998.

[42] E. Koski, “Concepts for a reliable multicast data link protocol for HF radio communications,” in *MILCOM 2005-2005 IEEE Military Communications Conference*, 2005, pp. 681-687: IEEE, 2005.

[43] S. Medhekar, “Roll-Call: an energy efficient radio frequency identification system,” Rutgers University-Graduate School-New Brunswick, 2007.

[44] D. Weissman, A. H. Levesque and R. A. Dean, “Interoperable wireless data,” *IEEE Communications Magazine*, vol. 31, no. 2, pp. 68-77, 1993.

[45] W. C. Lee, *Mobile communications engineering*. McGraw-Hill Professional, 1982.

[46] A. M. Michelson and A. H. Levesque, “Error-control techniques for digital communication,” *New York, Wiley-Interscience, 1985, 483 p.*, 1985.

[47] R. Sinha and S. Gupta, “Performance evaluation of a protocol for packet radio network in mobile computer communications,” *IEEE transactions on vehicular technology,* vol. 33, no. 3, pp. 250-258, 1984.

[48] U. Svasti-Xuto, Q. Wang and V. Bhargava, “Capacity of an FH-SSMA system in different fading environments,” *IEEE transactions on vehicular technology*, vol. 47, no. 1, pp. 75-83, 1998.

[49] S. L. Ariyavisitakul and G. M. Durant, “A broadband wireless packet technique based on coding, diversity, and equalization,” *IEEE Communications Magazine,* vol. 36, no. 7, pp. 110-115, 1998.

[50] A. Fiser, “Mapping the long-term options for Canada's north: Telecommunications and broadband connectivity,” 2013: Conference Board of Canada, 2013.

[51] M. Levitt, “Nation-building at home, vigilance beyond: Preparing for the coming decades in the arctic,” Speaker of the House of Commons, Ottawa, 2019.

Automatic Analysis of Nonverbal Mirroring Communication

Oky Dicky Ardiansyah Prima, Yuta Ono
Graduate School of Software and
Information Science, Iwate Pref. Univ.
Takizawa, Japan
email: prima@iwate-pu.ac.jp,
g236s001@s.iwate-pu.ac.jp

Kumiko Hosogoe, Miyu Nakano
Faculty of Social Welfare,
Iwate Prefectural University
Takizawa, Japan
email: hosogoe@iwate-pu.ac.jp,
g221r007@s.iwate-pu.ac.jp

Takashi Imabuchi
Office of Regional Collaboration,
Iwate Pref. Univ.
Takizawa, Japan
email: t_ima@ipu-office.iwate-pu.ac.jp

***Abstract*—Nonverbal communication plays an important role in social interaction. Mirroring, an action that mimics the nonverbal behavior patterns of their interaction partners, captures the attention of the Human-Computer Interaction (HCI) community. This action can help building rapport with others by making communication more effective and reflective. This study proposes a computer vision-based system that detects mirroring and analyzes the time lag during a face-to-face communication. Our approach consists of the following steps: (1) human pose estimation; (2) hand gestures quantization; (3) action detection based on Dynamic Time Warping (DTW); (4) estimation of mirroring time lag based on the cross-correlation. For this study, we recorded twenty face-to-face communication scenes using an omni-directional video camera with and without mirroring performed by the imitator. Results show that the DTW was able to detect actions having distinct gestures, whereas the cross-correlation was able to estimate the time lags for reactive mimicry of the imitator during the conversation.**

Keywords- mirroring communication; nonverbal communication; human pose estimation; DTW; cross-correlation.

I. Introduction

To improve communication skill, it is important to pay attention to eye contact, gestures, postures, body movements, and voice tones. These nonverbal actions can provide clues, additional information, and meaning in addition to verbal communication. Moreover, using these actions that reflect the behavior of the talking partner can help to create a strong connection with both side during the conversation. These techniques are called nonverbal mirroring. This study extends our previous research on analysis of communication mirroring using vision cameras [1]. Nonverbal mirroring during face-to-face communication can be used to show empathy and positive reaction to counterparts. Nonverbal behavioral mimicry can occur with little or no awareness but can occur during more than 30% of a given interaction [2]. More specifically, nonverbal mirroring can be distinguished from imitation behavior, which is an event in which two people act the same regardless of timing, and complementary behavior, which is an event in which two people act differently [3]. In this paper, we limited the scope of nonverbal mirroring to imitation behavior. The results of this study will complement studies on mirroring facial expressions in the generation of rapport scales.

The areas of the human brain that are activated by observation and execution of the same actions are called the "mirror neuron system." Functional Magnetic Resonance Imaging (fMRI) of frontal and parietal regions of the brain indicated that these regions are most consistently involved during mirroring [4]. For actions that are considered mirroring, each action taken by a partner are reciprocated by the coordinated manner with time lags. Studies have been conducted to define the time lag before mirroring occurs. Hale et al. (2020) suggested that 400-1,000ms is a plausible time range for reactive mirroring in a natural conversation [3]. However, mirroring might happen on a longer timescale, 2-10s [5] or 7s [6] at most.

Traditionally, measuring nonverbal mirroring had to be done manually by annotating of characteristic gestures of the subject and similar gestures of the counterpart from recorded videos of face-to-face communication. The resulting repetitive behavior, its duration, and response latency are quantized and used for further analysis, such as rapport-based behavior analysis [7]. BECO2, an integrated behavioral coding system, is widely used in Japanese universities to train students about behavior coding [8]. The system allows observers to record and analyze the occurrence and duration of actions by pressing the keyboard keys corresponding to each category. Because there is a lot of ambiguity in judgments of specific actions, observers tend to make inconsistent judgments, which reduces the quality of measurements.

The analysis of nonverbal mirroring has been studied for some time, but little research has been done on how to automate the analysis as an alternative to manual coding. To date, there is no practical software application program that can automatically detect mirroring from a video of face-to-face communication and calculate the time difference until the mirroring occurs. Speech and video processing technologies may contribute to the efficient analysis of conversational scenes. On the one hand, speech processing can reveal nonverbal behaviors such as speech, stress acts, and speech rate based on speech signals [9], but on the other hand, video processing can measure facial information and gestures [10].

Since mirroring is a complex phenomenon [11], we made a first attempt to build a framework which quantify gestures from a video of face-to-face communication in order to automatically detect the presence of mirroring [1]. In this paper, we further enhance the framework by improving

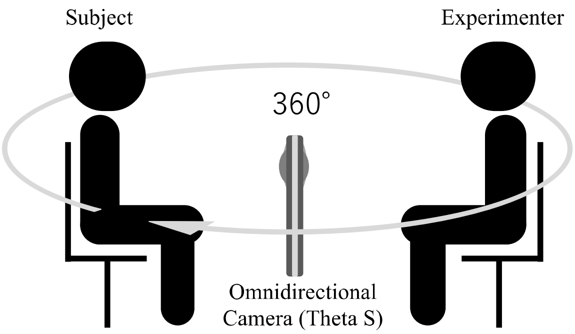


Figure 1. Experimental setting.

gesture detection and adding the ability to estimate the overall time lag of the detected mirroring.

The rest of this paper is organized as follows. Section II presents related research on behavioral coding and gesture analysis using computer vision techniques. Section III introduces the proposed framework for analyzing the presence of communication mirroring. Section IV describes the results. Finally, Section V presents our concluding remarks.

II. Related Work

The development of automatic mirroring detection involves building the mimicry dataset. MAHNOB is a public mimicry dataset consisting of a collection of multisensory audiovisual recordings of fully synchronized naturalistic dyadic interactions. The recordings were made under controlled laboratory conditions using 15 cameras and three microphones to obtain the most favorable conditions possible for analyzing the observed behavior [12]. Bilakhia et al. (2015) applied classifiers such as cross-correlation, generalized time warping, and Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to face and head movement data in the MAHNOB dataset [13]. However, the mirroring detection performance of these classifiers was poor, suggesting that more advanced learning methods are needed to deal with the variability in the dataset.

Some studies have attempted to detect mirroring using special cameras. Terven et al. (2015) introduced mirroring detection based on head-gestures using computer vision-based wearable devices [14]. A camera embedded in the wearable device was used to detect facial features of the partner during a face-to-face communication. Hidden Markov Models (HMMs) was used to recognize similar head-gestures. However, the mirroring detection performance was affected by the amount of head-gestures that occur in each data case. Jaana et al. (2014) developed an automated behavioral analysis system using a single omnidirectional camera. This system analyzed facial expressions, head nods, utterances based on facial features extracted from the camera [10]. While the system is not specifically designed to detect mirroring, it opens a way to simplify the video recording process during face-to-face communication by using an omnidirectional camera to analyze all participants in a conversation.

Body movements can be automatically accessed using computer vision-based methods, such as Motion Energy Analysis (MEA) [15] and OpenPose [16]. MEA measures motion by counting color changes in successive frames within a predefined region of interest, whereas OpenPose measures key points on the human body, hands, face and feet. Schoenherr et al. (2019) evaluated the performance of various time series analysis methods on nonverbal synchronous data quantified by MEA [17]. Schneider et al. (2019) proposed a gesture recognition system [18] using human posture obtained from a single camera using OpenPose. This system combined Dynamic Time Warping (DTW) and One-Nearest-Neighbor classifier to classify the time series data.

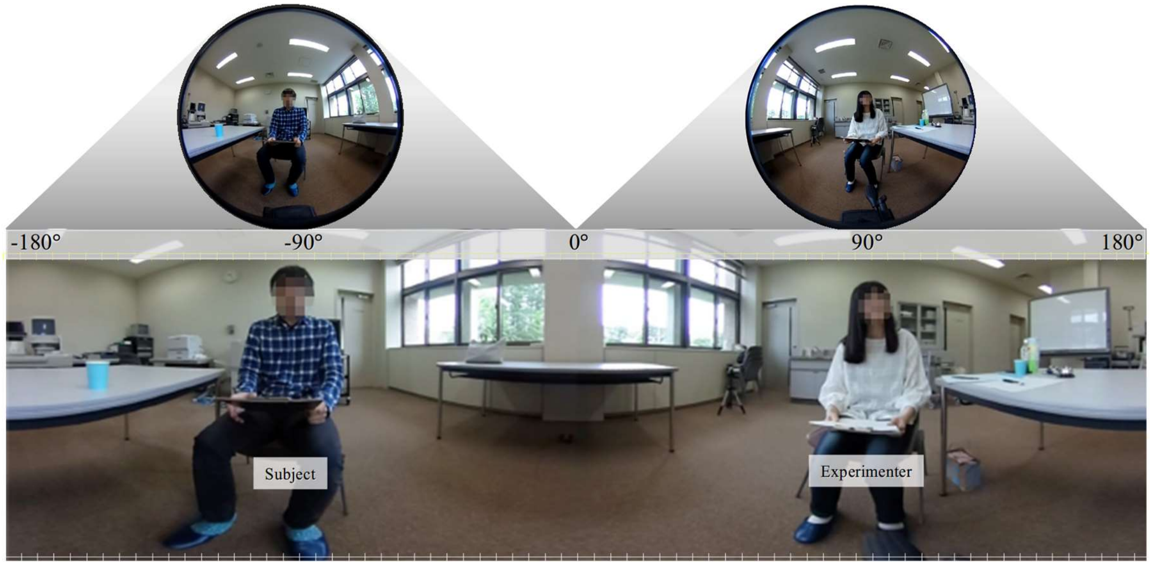


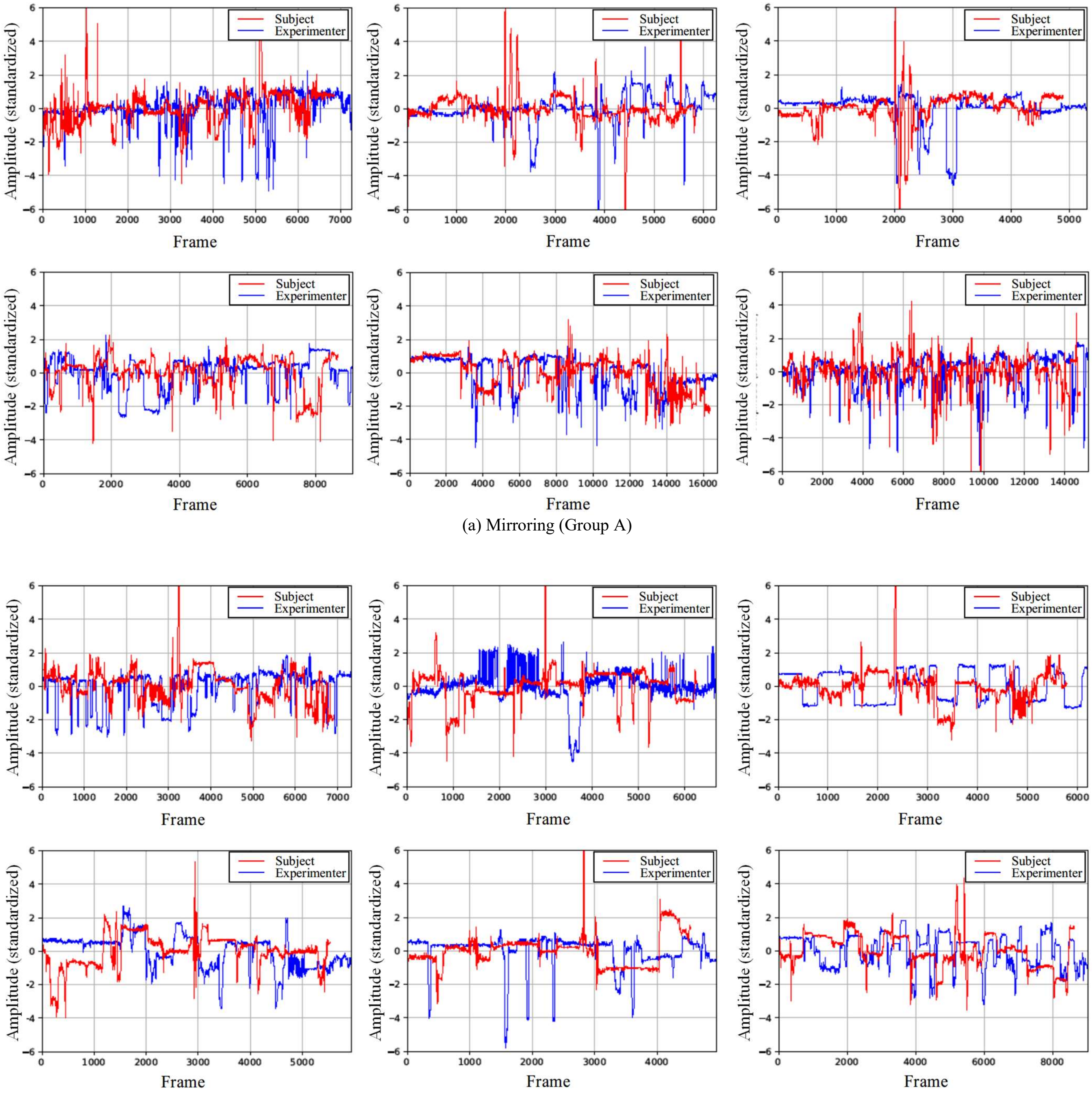
Figure 2. Face-to-face communication captured in a panoramic image

III. Analysis of Nonverbal Mirroring Communication

We propose a method to automatically analyze nonverbal mirroring communication from the recorded movements of pairs of participants (dyads): a subject and an experimenter. Our method uses DTW to detect and classify characteristic movements and uses cross-correlation to estimate the overall time lag of mirroring movements during a conversation. Here, we chose to focus on hand-gesture data only, based on a pilot study that revealed stronger similarities between the dyads than whole-body posture data.

A. Participants

46 students (25 males and 21 females) who have had part-time work experience in multiple faculties at Iwate Prefectural University were interviewed for approximately 5 minutes. The experimenter was a student in the same grade as the subjects and had been fully trained in behavioral mirroring. Subjects, who were not normally close to the experimenter, were recruited through the snowball sampling method.



(a) Mirroring (Group A)

(b) Without Mirroring (Group B)

Figure 3. An example of Wrist data for Group A and B.

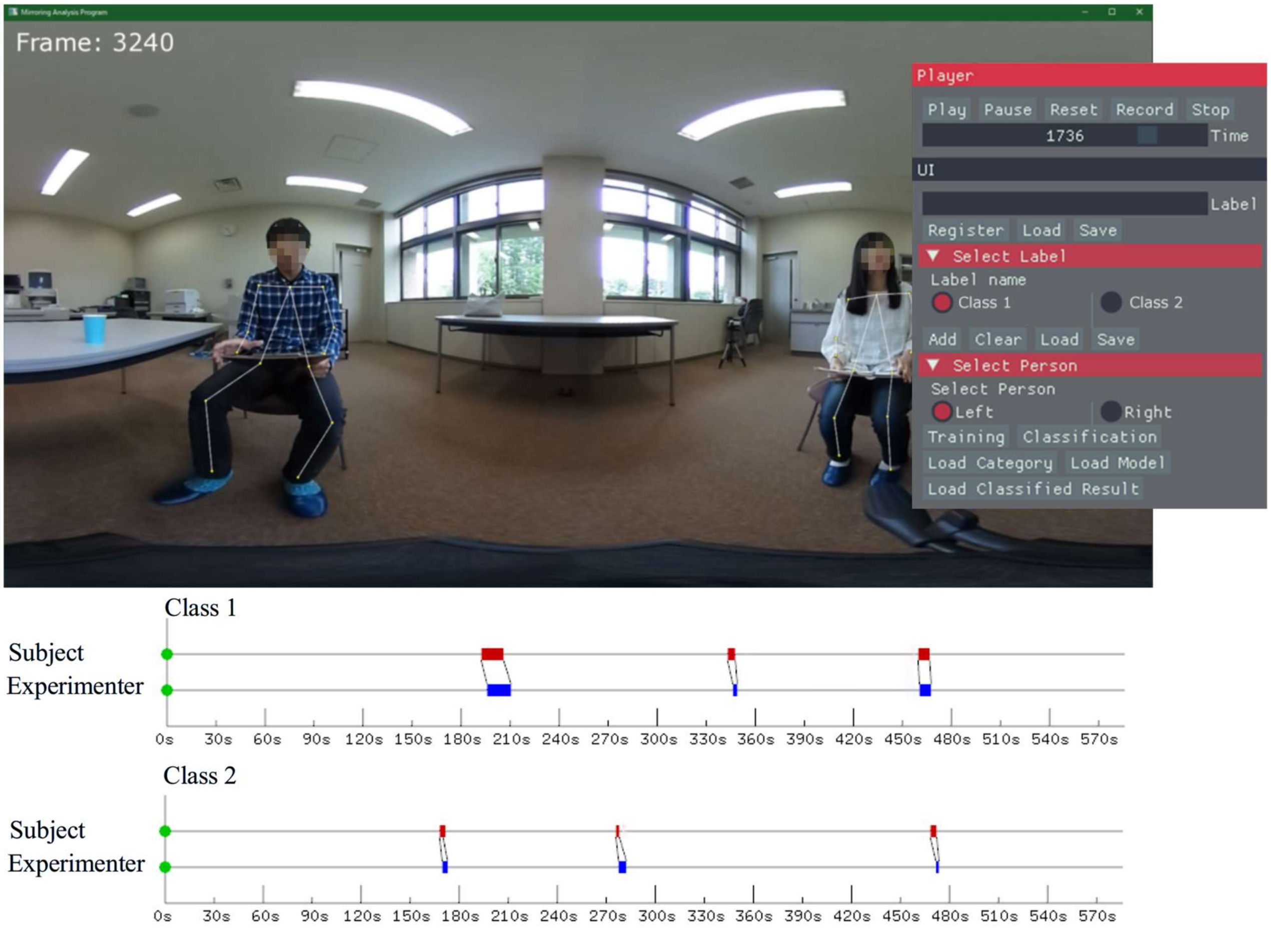


Figure 4. The Graphical User Interface (GUI) created to facilitate the selection and visualization of the training dataset in this study.

During the interview, the subjects were asked about their work experience. The experimenter intentionally mirrored 25 subjects (14 males and 11 females) and did not mirror the remaining 21 subjects (11 males and 10 females). Hereinafter, we refer to the former as Group A and the latter as Group B. This division was done randomly. Six subjects who did not perform the hand gestures were excluded from the analysis. Finally, a total of 40 conversations, 20 in group A (10 males and 10 females) and 20 in group B (11 males and 9 females), were included in the analysis.

B. *Laboratory Setup*

Two chairs were set up in the room facing each other for the subjects and the experimenter, as shown in Figure 1. To simplify the video recording process, an omnidirectional camera (Ricoh Theta S) [19] was placed between the experimenter and the subject, and video recording was performed at 30 Hz. Subjects were seated after completing the informed consent form. Subjects and experimenters were given a clipboard to take the necessary notes on. The interviews were conducted while holding this clipboard. This clipboard restricted the subject's hand gestures and allowed us to efficiently extract only those gestures that are important for communication mirroring. The subject was expected to generate hand gestures with one hand while holding the clipboard in the other hand, or to generate hand gestures with both hands by placing the clipboard on his or her lap.

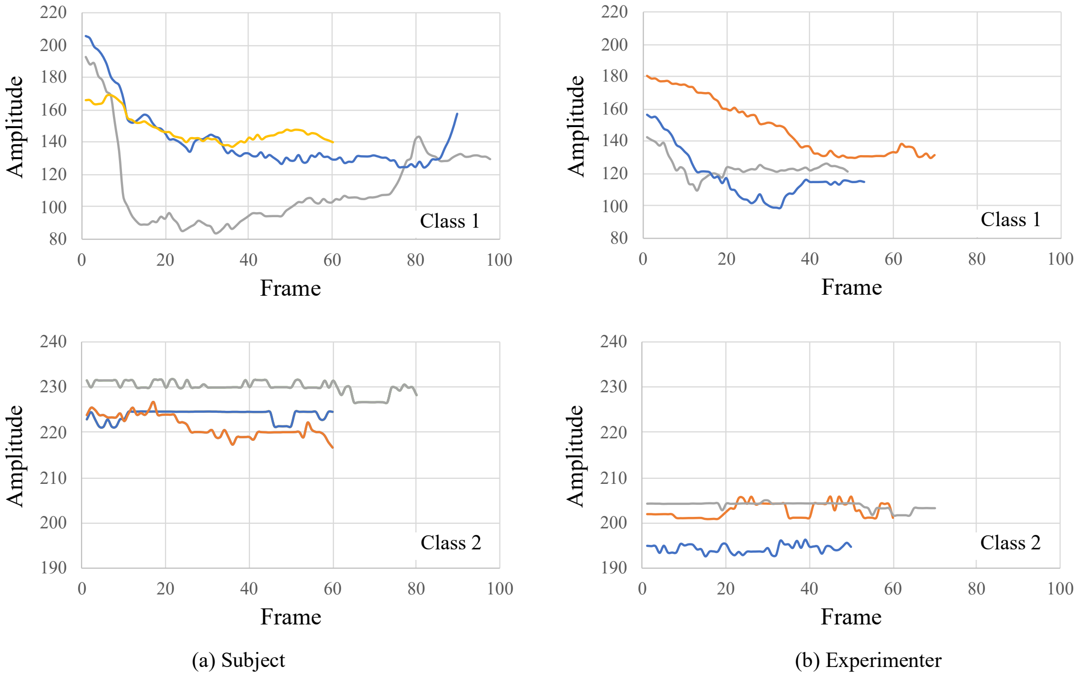
C. *Pre-processing*

1) Panoramic Image Projection

Ricoh Theta S generates two fisheye images to represent a 360° image. We merged and warped these images to produce a panoramic image, as shown in Figure 2, which allows the experimenter and the subject in the image to be seen from the front. The panoramic image is presented as a rectangular image of a 360° image. No special effort was made to produce this image. Everything was done using the built-in Ricoh utility program.

2) Quantitation of Hand Gestures

OpenPose with the 18-keypoint Coco body model was used to estimate the body posture of the dyads. For the purpose of this paper, only the positional information of the wrist joint is extracted from the body posture. The pixel coordinates of each joint were calculated from the panoramic images and these coordinates were normalized with the neck joint as the



(a) Subject (b) Experimenter

Figure 5. Two classes of gestures created from dyads for this study.

origin. We quantified the wrist data as follows. Let $W_n(x_l, y_l, x_r, y_r)$ represents coordinates of both wrists at n^{th} frame, wrist data at this frame is calculated by

$$Wrist_n = \|W_n\|_2 = \sqrt{x_l^2 + y_l^2 + x_r^2 + y_r^2} \ . \qquad (1)$$

Equation (1) reduces the dimensionality of the data at both wrists, but our preliminary results show that this data transformation can eliminate ambiguous gestures that do not correspond to mirroring.

In this study, we use wrist time series data to represent the gesture. Figure 3 shows an example of wrist data for Group A and B, measured in this study. These figures show that the variability of the dyads' wrist data was similar in Group A, but not in Group B. Here, we did not limit the duration of the interviews, which resulted in different frame lengths of the wrist data measured in each interview.

3) Building Training Dataset for DTW

Given that natural mirroring does not require a faithful reproduction of the opponent's hand gestures, we defined two classes of hand gestures based on their movement characteristics. We considered that highly variable gestures are subject to mirroring and less variable gestures are not. Hereinafter, we refer to them as class 1 and class 2, respectively.

A Graphical User Interface (GUI) was created to facilitate the selection and visualization of wrist data ranges for building training dataset, as shown in Figure 4. The slider can be used to determine the onset and offset of each gesture. In addition, when the slider is moved, the frame and the posture of both persons corresponding to the slider are displayed in real time. Training data individually selected from subjects and experimenters are drawn in a timeline. The vertical lines connecting the subject and the experimenter indicates that the subject's gestures are mirrored.

TABLE I. LIKELIHOOD BETWEEN GESTURES OF THE SUBJECT AND THE EXPERIMENTER.

NO.	CLASS LABEL (SUBJECT)	PREDICTED CLASS LABEL (EXPERIMENTER)	MAXIMUM LIKELIHOOD
1.	1	1	0.599
2.	1	1	0.714
3.	1	1	0.716
4.	2	2	0.912
5.	2	2	0.765
6.	2	2	0.766

One of the most difficult aspects of creating a training dataset is to define the number of classes of gestures in the dataset. After reviewing all recorded videos, none of the subjects generated two-handed gestures with the clipboard on their lap. When generating a gesture, the subject always holds the clipboard with one hand. Interestingly, when the gesture was not generated, the subject continued to hold the clipboard with both hands. Based on these observations, we classified the subjects' gestures into two classes.

Figure 5 shows the wrist data, describing the class 1 and class 2 hand gestures created from the dyads, respectively. Each class contains three time series data. In class 1, the

TABLE II. Results of cross-correlation analysis for 20 mirrored conversations in this study

No.	Time Lags		Maximum Correlation	Critical Value
	Frame	Time (s)		
1.	-79	-2.6	0.476	0.027
2.	-60	-2.0	0.725	0.021
3.	-47	-1.6	0.287	0.029
4.	-67	-2.2	0.433	0.030
5.	-68	-2.3	0.174	0.025
6.	-38	-1.3	0.419	0.023
7.	-147	-4.9	0.715	0.020
8.	-83	-2.8	0.301	0.033
9.	-72	-2.4	0.298	0.029
10.	-97	-3.2	0.489	0.018
11.	-110	-3.7	0.731	0.016
12.	-62	-2.1	0.737	0.023
13.	-21	-0.7	0.315	0.022
14.	-55	-1.8	0.877	0.033
15.	-49	-1.6	0.669	0.017
16.	*-481*	*-16.0*	*0.639*	*0.021*
17.	*-812*	*-27.1*	*0.257*	*0.032*
18.	*933*	*31.1*	*0.312*	*0.030*
19.	*433*	*14.4*	*0.156*	*0.021*
20.	*1,279*	*42.6*	*0.289*	*0.031*

TABLE III. Results of cross-correlation analysis for 20 non-mirrored conversations in this study

No.	Time Lags		Maximum Correlation	Critical Value
	Frame	Time (s)		
1.	12,271	409.0	0.284	0.050
2.	-2,736	-91.2	0.215	0.037
3.	475	15.8	0.686	0.040
4.	1,483	49.4	0.425	0.033
5.	689	23.0	0.241	0.032
6.	219	7.3	0.203	0.032
7.	230	7.7	0.340	0.024
8.	-751	-25.0	0.245	0.033
9.	41	1.4	0.513	0.035
10.	-1,802	-60.1	0.335	0.041
11.	1,505	50.2	0.257	0.030
12.	662	22.1	0.244	0.021
13.	1,032	34.4	0.197	0.023
14.	-2,027	-67.6	0.147	0.034
15.	2,777	92.6	0.160	0.024
16.	1,025	34.2	0.235	0.025
17.	4,210	140.3	0.265	0.077
18.	*-5*	*-0.2*	*0.325*	*0.027*
19.	*-107*	*-3.6*	*0.474*	*0.035*
20.	*-219*	*-7.3*	*0.244*	*0.027*

moment the hand leaves the clipboard was determined to be onset, and the moment the hand returns to the clipboard is to be offset. On the other hand, in the class 2, the onset and offset were determined when the clipboard is held in both hands for a period.

The gestures of the experimenter are relatively shorter than those of the subjects. This can be interpreted as a result of the experimenter confirming the subject's gesture and then simply imitating the gesture.

4) Detection

We performed DTW using the Gesture Recognition Toolkit (GRT) of Gillian and Paradiso (2012) [20] and applied maximum likelihood from the warping distance to estimate similarity to the training data. Table I shows the likelihood between gestures of the dyads shown in Figure 5. This data was normalized before it was inputted into the GRT. The DTW was able to correctly match the same gesture between the dyads, even though the length of the subject's gesture is different from the length of the experimenter's gesture. Although more gesture training data would be desirable, this study focuses on a basic analysis of the extent to which the simplest gestures can be used to detect what appears to be mirroring.

In this study, we used the average length of the training data frame as the maximum warp amount during the calculation of DTW. We considered that this DTW's window width is sufficient for our purpose.

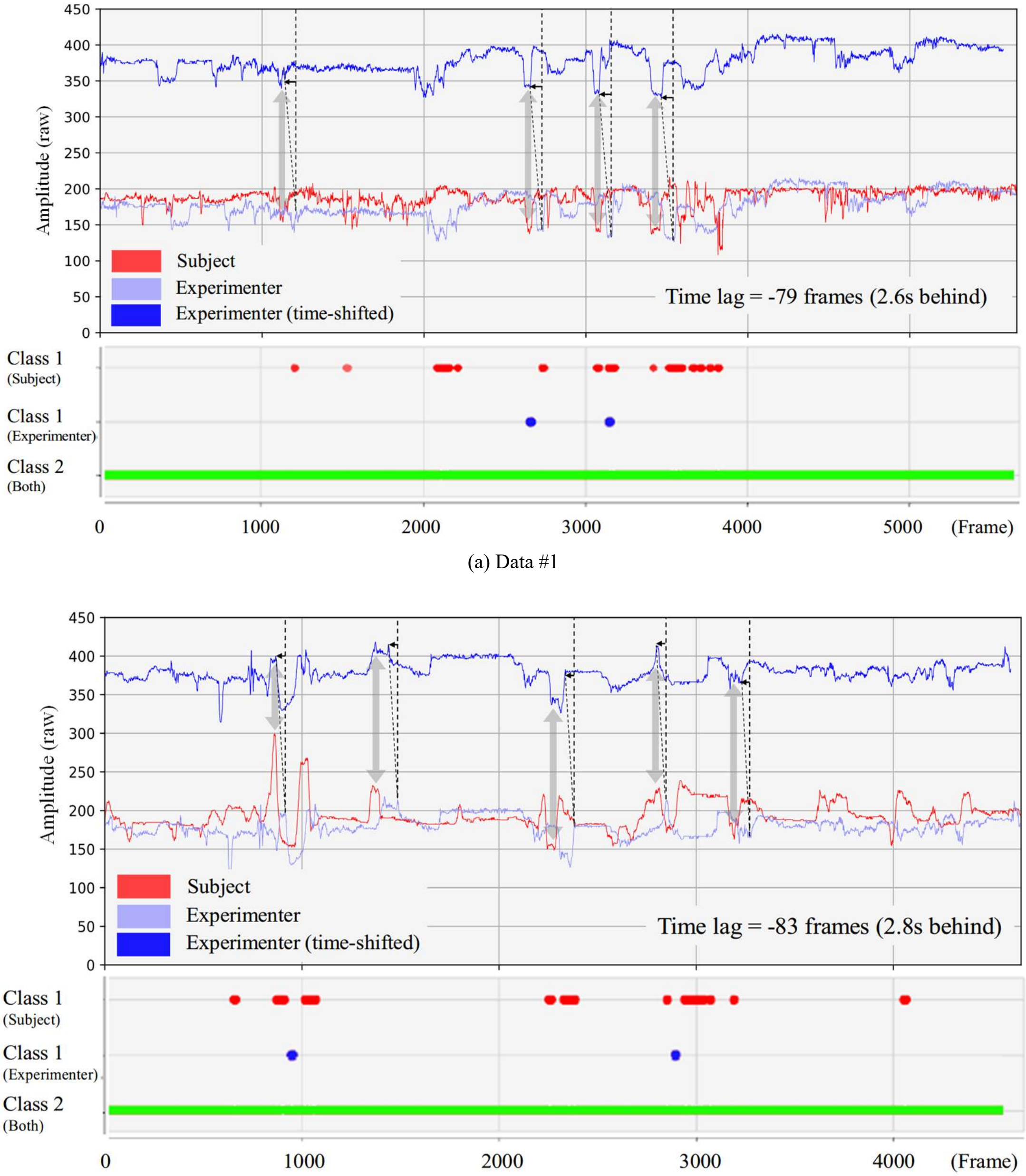
5) Cross-correlation

*C*ross-correlation is useful for aligning two time series, one of which is lagged relative to the other, since its peak occurs at the lag where the two time series are most correlated. Cross-correlation ρ at delay d between the subject's and the examiner's hand gestures is define as

$$\rho(d) = \frac{\sum_{i=1}^{N}\{(x_i - \bar{x})(y_{i-d} - \bar{y})\}}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \ \sum_{i=1}^{N}(y_{i-d} - \bar{y})^2}} \qquad (2)$$

Here, x_i and y_{i-d} are series of the subject's and the examiner's hand gestures, respectively. $\bar{x}$ and $\bar{y}$ represents means of x_i and y_{i-d}.

Following [3] and [5], we assume that the mirroring occurs at around 0.4-7s. In other words, if a time lag longer than 7s is calculated, it means that no mirroring has occurred in the data. From this perspective, cross-correlation can be the easiest way to determine the presence or absence of mirroring.



(a) Data #1

(b) Data #8

Figure 6. Detection and cross-correlation of the two classes in the two mirrored conversations.

IV. Results

A. Mirroring Analysis

For automatic detection of two classes, we used only training data collected from subjects, as shown in Figure 5(a). The experimenter's gestures described in the previous section were only for validating the subject's gestures. The behaviors of subjects and experimenters corresponding to each class were collected and presented in a time series of the input data.

Cross-correlations were calculated for the subject's and experimenter's wrist data. Here, we normalized the values of the cross-correlation to take values from -1 to +1. The maximum value of the cross-correlation function indicates the point in time where the data are most aligned (delay time).

Figure 6 shows the results of the detection and cross-correlation of the two classes in the two interview scenes. The top of the figure shows the raw data of the dyads' wrist movements. The experimenter's data were then shifted based on the time lag between dyads as measured by cross-correlation. To make the shifted data easier to observe, the shifted data was moved upward. The gray arrows indicate some of the areas where the subject's wrist data and the shifted experimenter's wrist data are similar. The bottom row of the figure shows two classes of detection from the dyads' wrist

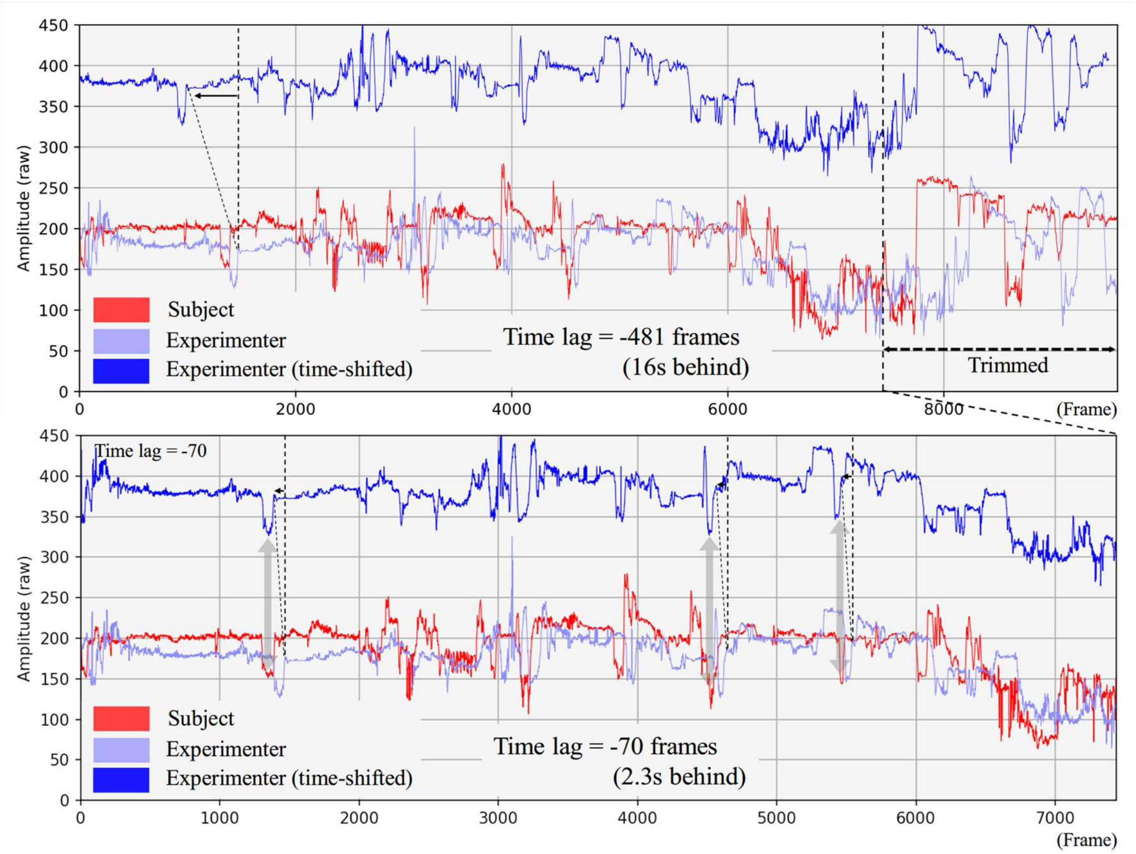


Figure 7. Refinement of the cross-correlation result by data trimming (data #16).

data. Red points represent class 1 detections for the subject, whereas blue points for the experimenter. Given that these detected gestures have similar features, it is likely that these blue dots mirror the red dots. As can be seen from the figure, the number of class 1 detections of the subjects was higher than that of the experimenters. It is a reasonable result since the incidence of the mirrored gestures should not exceed the mirrored target. It should also be noted that the class 1 behaviors detected from the experimenter occurred after the same class of behaviors detected from the subjects. Green dots represent class 2 detections of the dyads. Gestures that could not be classified as class 1 were classified as class 2, resulting in a higher frequency of class 2 detections. In this mirroring analysis, we did not target class 2, which has few variable gestures, so we integrated the resulting detections of class 2 of the dyads, as shown in Figure 6.

Table II shows results of cross-correlation analysis for 20 mirrored conversations in this study. In the first 15 conversations, we found that the experimenter's wrist data are delayed between 0.7-4.9s than the subject's wrist data. However, the remaining five conversations measured a time lag that did not meet the mirroring condition (numbers in italic). These are partly due to the inability of the experimenter to mirror the subject properly during conversations, but also due to the increased variation in behavior, such as having a drink during a conversation. Figure 7 shows that the measurement delay of conversation #16 can be corrected from 481 frames (16s) to 70 frames (2.3s) by trimming a portion of the data.

Table III shows results of cross-correlation analysis for 20 non-mirrored conversations in this study. In the first 17 conversations, the measured time lag showed that no mirroring occurred during the conversation. However, the time lag in the remaining three conversation scenes suggests that mirroring behavior occurred (numbers in italic). Figure 8 shows that similar patterns between the dyads in conversation #18 and #19. This implies that the experimenter may unconsciously engage in mirroring behavior.

B. *Perceived Empathy*

Perceived empathy was measured using the 16-item empathy understanding subscale of the Barrett-Lennard Relationship Inventory (BLRI) [21]. All items in the perceived empathy assessment were scored on a six-point Likert scale, with 1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = slightly agree, 5 = agree, and 6 = strongly agree. Higher scores represented more empathy perceived by subjects. Cronbach’s coefficient alpha is 0.81, indicating high reliability.

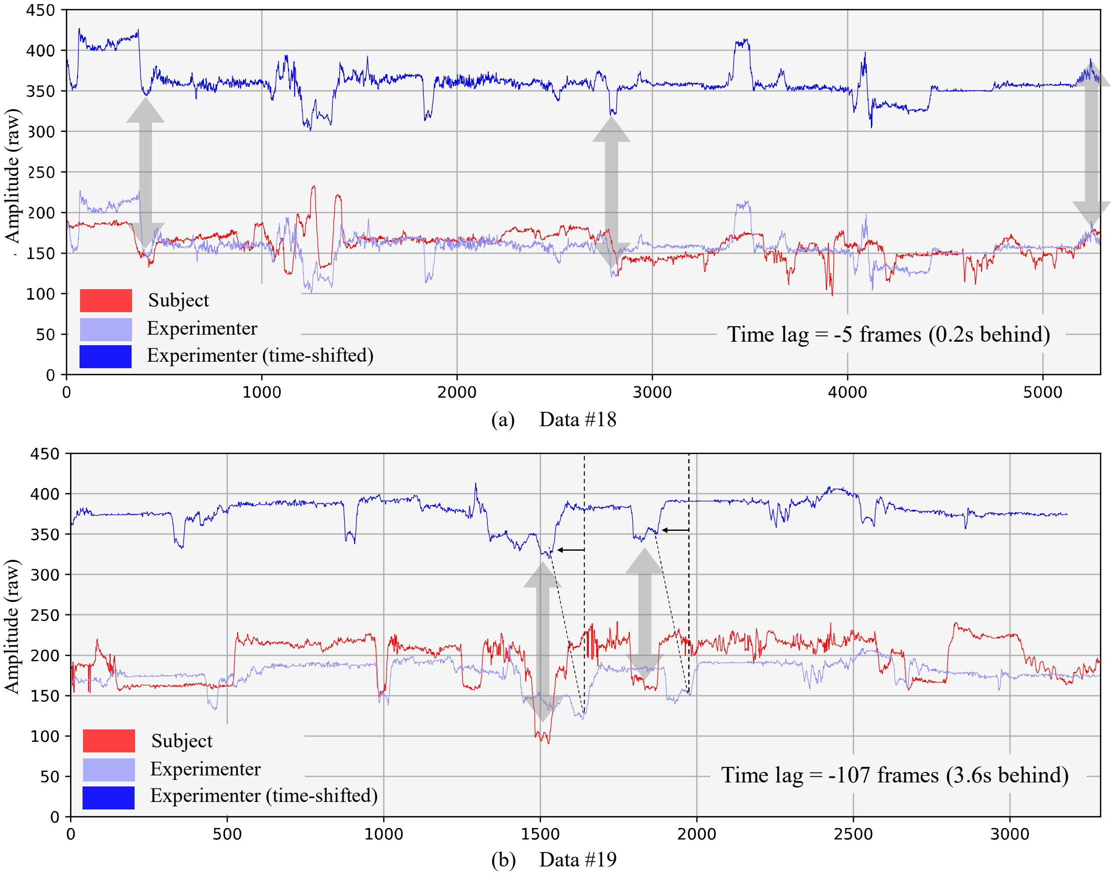


Figure 8. Similar patterns between the dyads in the non-mirroring conversations.

TABLE IV. MEAN AND STANDARD DEVIATION OF THE PERCEIVED EMPATHY SCORES

GENDER	WITH MIRRORING	WITHOUT MIRRORING
MALE	4.37 (0.52)	4.23 (0.49)
FEMALE	4.77 (0.48)	4.65 (0.37)

To evaluate whether there was a difference in the mean score on the perceived empathy scale by gender and the presence of mirroring, we conducted a two-factor analysis of variance. The means and standard deviations are shown in Table IV. The results showed a significant main effect on the gender factor ($F(1, 36) = 8.04$, $p < 0.5$), but not on the presence of mirroring factor ($F(1, 36) = 0.72$, *n.s.*). No significant interactions were observed ($F(1, 36) = 0.01$, *n.s.*). The present results indicate that female subjects were more emphatic with and without mirroring during the interview. To promote empathy by mirroring communication, a longer interview than the present experiment would be necessary.

C. Debriefing

Once the interview was completed, the experimenter provided the subject with accurate and pertinent information about the nature of this experiment. During the debriefing process, subjects are informed about what the hypothesis for the experiment was as well.

After the debriefing, the 20 subjects who had been mirrored were asked if they noticed that they were being mirrored or if they felt that the conversation was unnatural. Six subjects said that they noticed that they were being mirrored. They were aware of being mirrored because they were already knowledgeable about mirroring. All 20 subjects did not find it unnatural to be mirrored. In this short interview, the recorded video shows that there is almost no pause in the conversation between the two parties, with the subject actively answering the experimenter's questions. This may explain why the subjects did not feel unnatural in their conversations.

V. CONCLUDING REMARKS

Many studies have been done to automatically detect mirroring during conversations, but to the best of our knowledge, it has not reached a practical level [13][14][22]. In order to automatically detect mirroring behavior during conversations, we attempted to make the gestures on the dyads as simple as possible. Having the subjects hold the clipboard during the conversation was effective in limiting their hand gestures. This allowed the experimenter to properly imitate the subjects' gestures.

The DTW and cross-correlations used in this study are commonly used in time series data analysis. In general, a mimic gesture is one in which the performer tries to mimic the actions of the other person as accurately as possible. However, there is a difference between similar behaviors perceived by humans and similar behaviors seen from sensor information. For this reason, we successfully identified distinct gestures using DTW without being too obsessed with small movements by making the information of the hand gestures one-dimensional using the L2 norm. On the other hand, cross-correlation analysis successfully estimated the time lag of mirroring behavior in conversations. Interestingly, cross-correlation analysis was able to detect the unintended mirroring even if the experimenter did not intend to mirror the subject.

Nonverbal mirroring communication helps to create a strong connection between the two parties during a conversation, but in the present experiment, subjects' empathy levels were not significant with or without mirroring communication. A longer interview than the present experiment would be necessary to promote empathy through mirrored communication.

Finally, since the observed data contain a lot of noise, removing the noise can improve the accuracy of the analysis. The mirroring time lag can be properly determined by trimming a portion of the data, as in this study.

Acknowledgment

We thank the faculty of Social Welfare, Iwate Prefectural University, Japan, for funding this project. We also thank the editor and four anonymous reviewers for their constructive comments, which helped us to improve the manuscript.

References

[1] K. Hosogoe, M. Nakano, O. D. A. Prima, and Y. Ono, “Toward automated analysis of communication mirroring,” The Thirteenth International Conference on Advances in Computer-Human Interactions, ACHI2020, pp. 15-18, 2020.

[2] J. L. Lakin and T. L. Chartrand, “Using nonconscious behavioral mimicry to create affiliation and rapport,” Psychological Science, 14(4), pp. 334–339, 2003.

[3] J. Hale et al., “Are you on my wavelength? Interpersonal coordination in dyadic conversations,” Journal of Nonverbal Behavior, 44 (1), pp. 63-83, 2020.

[4] P. Molenberghs, R. Cunnington, and J. B.Mattingley, “Is the mirror neuron system involved in imitation? A short review and meta-analysis,” Neuroscience and Biobehavioral Reviews, 33 (7), pp. 975-980, 2009.

[5] N. P. Leander, T. L. Chartrand, and J. A. Bargh, “You give me the chills: Embodied reactions to inappropriate amounts of behavioral mimicry,” Psychological Science, 23(7), pp. 772–779, 2012.

[6] J. W. Robinson, A. Herman, and B. J. Kaplan, “Autonomic responses correlate with counselor–client empathy.” Journal of Counseling Psychology, 29(2), pp. 195–198, 1982.

[7] C. F. Sharpley, J. Halat, T. Rabinowicz, B. Weiland, and J. Stafford, “Standard posture, postural mirroring and client-perceived rapport.” Counselling Psychology Quarterly, 14(4), pp. 267–280, 2001.

[8] Behavior coding system, DKH Co. Ltd., https://www.dkh.co.jp/product/behavior_coding_system/ [retrieved: August 31, 2020]

[9] K. Otsuka and S. Araki, “Audio-visual technology for conversation scene analysis,” NTT Technical Review, 7(2), pp. 1-9, 2009.

[10] Y. Jaana, O. D. A. Prima, T. Imabuchi, H. Ito, and K. Hosogoe, “The development of automated behavior analysis software,” Proc. SPIE 9443, Sixth International Conference on Graphic and Image Processing (ICGIP), pp. 1-5, 2014.

[11] T. L. Chartrand and J. A. Bargh, “The chameleon effect: the perception-behavior link and social interaction,” Journal of Personality and Social Psychology, 76(6), pp. 893–910, 1999.

[12] MHI-Mimicry database, https://mahnob-db.eu/mimicry/ [retrieved: August 31, 2020]

[13] S. Bilakhia, S. Petridis, A. Nijholt, and M. Pantic, “The MAHNOB mimicry database: a database of naturalistic human interactions,” Pattern Recognition Letters, 66, pp. 52–61, 2015.

[14] J. R. Terven, B. Raducanu, M. E. Meza-de-Luna, and J. Salas, “Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices,” Neurocomputing, 175, pp. 866–876, 2015.

[15] K. Grammer, M. Honda, A. Juette, and A. Schmitt, “Fuzziness of nonverbal courtship communication unblurred by motion energy detection,” Journal of Personality and Social Psychology, 77 (3), pp. 487-508, 1999.

[16] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” Computer Vision and Pattern Recognition, pp. 7291-7299, 2017.

[17] D. Schoenherr et al., “Quantification of nonverbal synchrony using linear time series analysis methods: Lack of convergent validity and evidence for facets of synchrony,” Behavior Research Methods, 51(1), pp. 361–383, 2019.

[18] P. Schneider, R. Memmesheimer, I. Kramer, and D. Paulus, “Gesture recognition in RGB videos using human body keypoints and dynamic time warping,” Lecture Notes in Computer Science, vol. 11531, pp. 281–293, 2019.

[19] Ricoh Theta S, https://theta360.com/en/about/theta/s.html [retrieved: August 31, 2020]

[20] N. Gillian and J. A. Paradiso, “The gesture recognition toolkit,” Journal of Machine Learning Research, 15, pp. 3483–3487, 2014.

[21] G. T. Barret-Lennard, “Dimensions of therapiat responses as causal factors in therapeutic change,” Psycological Monographs, 76 (43), pp. 1-36, 1962.

[22] S. Michelet, K. Karp, E. Delaherche, C. Achard, and M. Chetouani, “Automatic imitation assessment in interaction,” Lecture Notes in Computer Science (Included in Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7559 LNCS, pp. 161–173, 2012.