PROCEEDINGS

of the

2018 Symposium on Information Theory and Signal Processing in the Benelux

May 31-1 June, 2018, University of Twente, Enschede, The Netherlands

https://www.utwente.nl/en/eemcs/sitb2018/

Luuk Spreeuwers & Jasper Goseling (Editors)

ISBN 978-90-365-4570-9

The symposium is organized under the auspices of

Werkgemeenschap Informatie- en Communicatietheorie (WIC)

& IEEE Benelux Signal Processing Chapter

and supported by

Gauss Foundation (sponsoring best student paper award)

IEEE Benelux Information Theory Chapter

IEEE Benelux Signal Processing Chapter

Werkgemeenschap Informatie- en Communicatietheorie (WIC)

UNIVERSITY OF TWENTE.

IEEE

Previous Symposia

1.	1980	Zoetermeer, The Netherlands, Delft University of Technology	
2.	1981	Zoetermeer, The Netherlands, Delft University of Technology	
3.	1982	Zoetermeer, The Netherlands, Delft University of Technology	
4.	1983	Haasrode, Belgium	ISBN 90-334-0690-X
5.	1984	Aalten, The Netherlands	ISBN 90-71048-01-2
6.	1985	Mierlo, The Netherlands	ISBN 90-71048-02-0
7.	1986	Noordwijkerhout, The Netherlands	ISBN 90-6275-272-1
8.	1987	Deventer, The Netherlands	ISBN 90-71048-03-9
9.	1988	Mierlo, The Netherlands	ISBN 90-71048-04-7
10.	1989	Houthalen, Belgium	ISBN 90-71048-05-5
11.	1990	Noordwijkerhout, The Netherlands	ISBN 90-71048-06-3
12.	1991	Veldhoven, The Netherlands	ISBN 90-71048-07-1
13.	1992	Enschede, The Netherlands	ISBN 90-71048-08-X
14.	1993	Veldhoven, The Netherlands	ISBN 90-71048-09-8
15.	1994	Louvain-la-Neuve, Belgium	ISBN 90-71048-10-1
16.	1995	Nieuwerkerk a/d IJssel, The Netherlands	ISBN 90-71048-11-X
17.	1996	Enschede, The Netherlands	ISBN 90-365-0812-6
18.	1997	Veldhoven, The Netherlands	ISBN 90-71048-12-8
19.	1998	Veldhoven, The Netherlands	ISBN 90-71048-13-6
20.	1999	Haasrode, Belgium	ISBN 90-71048-14-4
21.	2000	Wassenaar, The Netherlands	ISBN 90-71048-15-2
22.	2001	Enschede, The Netherlands	ISBN 90-365-1598-X
23.	2002	Louvain-la-Neuve, Belgium	ISBN 90-71048-16-0
24.	2003	Veldhoven, The Netherlands	ISBN 90-71048-18-7
25.	2004	Kerkrade, The Netherlands	ISBN 90-71048-20-9
26.	2005	Brussels, Belgium	ISBN 90-71048-21-7
27.	2006	Noordwijk, The Netherlands	ISBN 978-90-71048-22-7
28.	2007	Enschede, The Netherlands	ISBN 978-90-365-2509-1
29.	2008	Leuven, Belgium	ISBN 978-90-9023135-8
30.	2009	Eindhoven, The Netherlands	ISBN 978-90-386-1852-4
31.	2010	Rotterdam, The Netherlands	ISBN 978-90-710-4823-4
32.	2011	Brussels, Belgium	ISBN 978-90-817-2190-5
33.	2012	Enschede, The Netherlands	ISBN 978-90-365-3383-6
34.	2013	Leuven, Belgium	ISBN 978-90-365-0000-5
35.	2014	Eindhoven, The Netherlands	ISBN 978-90-386-3646-7
36.	2015	Brussels, Belgium	ISBN 978-2-8052-0277-3
37.	2016	Louvain-la-Neuve, Belgium	ISBN 978-2-9601884-0-0
38.	2017	Delft, The Netherlands	ISBN 978-94-6186-811-4

Preface

This event is the 39th edition of a sequence of annual symposia, that started in the 1980's, under the auspices of the Werkgemeenschap voor Informatie- en Communicatietheorie (WIC). Since 2011, the symposia are co-organized with the IEEE Benelux Signal Processing Chapter. The fruitfulness of this cooperation is also reflected by one common name:

"The 2018 Symposium on Information Theory and Signal Processing in the Benelux"

This year's venue is hotel De Broeierd in Enschede near the University of Twente. We are very fortunate to have two eminent keynote lecturers: Marcel Worring, director of the Informatics Institute of the University of Amsterdam and Petar Popovski, professor of Wireless Communications at Aalborg University in Denmark. Furthermore, there are 35 contributions, mostly by researchers from various universities in the Benelux countries, but also from companies and from universities elsewhere. These are all documented in the proceedings, either as an abstract or as a full paper, and presented at the symposium, either orally or via a poster. The social part of the symposium is a guided tour through the Grolsch Beer Factory in Enschede, which is also the site of the conference dinner.

We thank the keynote lecturers for accepting our invitation, all authors for their contributions to the scientific program, all participants for their presence, the Gauss Foundation for sponsoring the best student paper award.

Enschede, May 2018,

Luuk Spreeuwers and Jasper Goseling
(symposium organizers and proceedings editors)

Contents

Note: Except for the keynotes, papers are ordered after the last name of the first author

Keynote 1

Deep Learning for Multimedia

Marcel Worring

Abstract

With the advent of deep learning many applications of machine learning have surfaced ranging from medical diagnostics, autonomous driving, and product recommendation to social media analytics. For image analysis a lot of research has focused on classification of data using convolutional neural nets which are reaching very high performance. Recently graph based convolutional nets have started to gain momentum which focus on relations between different elements. We are extending these techniques to multimedia collections. In this presentation we show our work in this area where we first focus on techniques which are based on relations between items and from there we move on to techniques based on hypergraphs which can model relations and groups of items at the same time. Results of these techniques are shown on webforum data containing images, text, and metadata.

Biography

Prof. dr. Marcel Worring is a leading expert in multimedia research. He currently is the director of the Informatics Institute of the University of Amsterdam having over 200 fte on the broad range of computer science with among others top groups in computer vision, information retrieval and machine learning. As associate professor in the Informatics Institute, he performs research on multimedia analytics, bringing together image analysis, text analysis, machine learning and visualization. He also has an appointment as full professor in the Amsterdam Business School where his research is focused on applying state-of-the-art data science techniques in business applications. He is one of the founders of the recently launched Innovation Center for Artificial Intelligence. He was associate editor of IEEE Transactions on Multimedia and currently, is associate editor of ACM Transactions on Multimedia, and was general chair of the ACM Multimedia conference in 2016, the two leading journals and leading conference in the field. He has been leading and is participating in several national and European projects such as SortItOut (visual analytics for multimedia), VoxPol (fighting radicalization on the Internet), ASGARD (tools for Law Enforcement), VISTORY (visual Analytics for art history), and JOLT (digital journalism).

Keynote 2

How Reliability, Latency, Massiveness, and Blockchain are Transforming IoT Communication

Petar Popovski

Abstract:

The future wireless landscape, often associated with 5G, envisions three types of connectivity: enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communication (URLLC), and massive Machine Type Communication (mMTC). The latter two are seen as two generic types that support Internet of Things (IoT) communication, putting forward new types of requirements and research challenges, such as: protocols that operate with short packets, access for massive number of devices, techniques to achieve and assess extremely high reliability, etc. This set of challenges is further enriched by the advent of blockchain systems and smart contracts that allow autonomous interaction among IoT devices. The consensus protocols that set the basis for blockchain systems are critically reliant on communication, but they change the traffic pattern that has been envisioned for pre-blockchain IoT communication systems. This talk will give a perspective on the communication engineering challenges related to the emerging IoT communication systems, outline methods and architectures to solve them and provide communication-theoretic insights in some of the fundamental tradeoffs.

Biography:

Petar Popovski is a Professor of Wireless Communications with Aalborg University in Denmark. He received his Dipl. Ing and Magister Ing. degrees in communication engineering from the "Sts. Cyril and Methodius" in Republic of Macedonia, and the Ph.D. degree from Aalborg University in 2005. He is a Fellow of IEEE, a holder of a Consolidator Grant from the European Research Council (ERC), recipient of the Danish Elite Researcher Award, and a member of the Danish Academy for technical sciences (ATV). He is currently an Area Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, General Chair for IEEE SmartGridComm 2018 and General Chair for IEEE Communication Theory Workshop 2018. His research interests are in the area of wireless communication and networking, and communication theory.

Estimating Source-to-Electrode Transfer Functions in Atrial Electrograms

Bahareh Abdi[†], *Richard C. Hendriks*[†], *Alle-Jan van der Veen*[†], *and Natasja M.S. de Groot*[⋆]

[†]Circuits and Systems (CAS) Group, Delft University of Technology, the Netherlands
[⋆]Department of Cardiology, Erasmus University Medical Center, the Netherlands
[†]{b.abdikivanani, r.c.hendriks, a.j.vanderveen}@tudelft.nl
[⋆]{n.m.s.degroot}@erasmusmc.nl

Abstract

Atrial fibrillation (AF) is a common age-related cardiac arrhythmia characterized by rapid and irregular electrical activity of the atria. Persistence of AF is rooted in the electropathology of atrial tissue. As shown in previous studies (Yaksh 2015), analyzing electrogram signals recorded during intra-operative and high-resolution mapping of the entire atria can help to localize and quantify the degree of electropathology. These analyses can potentially be used to improve AF treatments and its earlier recognition which is currently hampered by the lack of appropriate signal processing methods that can be linked to the complex electrophysiological model. Moreover, some AF therapies like ablation of locations producing complex fractionated atrial electrograms (CFAEs) totally depend on the employed signal processing methods whose effectiveness is not yet clear (L. van der Does 2017).

There are many studies that employ well-known pure signal processing techniques to analyze atrial electrograms. These approaches include template matching of AF electrograms for the diagnosis of the electropathology (R. Houben 2006), evaluation of Shannon's entropy and the Kolmogorov-Smirnov (K-S) test as a measurement of signal complexity (Ng. Jason 2010), using sinusoidal recomposition and Hilbert transformations to characterize wave propagation and detect phase singularities (P. Kuklik 2015). Although these techniques offer discriminative features for classification of fractionated electrograms, they do not provide any insight on the underlying electrophysiological properties connected to these observations.

Our hypothesis is that understanding atrial fibrillation and improving AF therapies, starts with developing a proper model that is accurate enough (from a physiological point of view) and simultaneously simple enough to be used in a signal processing context. Understanding the problem of atrial fibrillation therefore starts with developing a proper, but simple, signal processing model that explains how signals propagates in tissue. We start this paper by presenting a linearized matrix representation of the well known reaction-diffusion equation governing the electrical propagation in atrial tissue. Using this representation we then propose a simplified electrogram model. One interesting application of this simplified model, is the deduction of the source-to-electrode transfer function. We show (by simulation) that these transfer functions can be used to describe the tissue's effect on the signal morphology, which is usually harder to obtain from the raw electrograms. Moreover, using simulations, we show that as long as the conductivity of the tissue and depolarization wavefront propagation in tissue are rather smooth, the transfer function consists of three extrema; two maximums and one minimum. The locations (in time) and amplitudes of these extrema with respect to each other, provides a summarized view of the electrical propagation and can be used as electro physiologically meaningful features for analysis of atrial activity and classification of CFAEs (Figure 1).

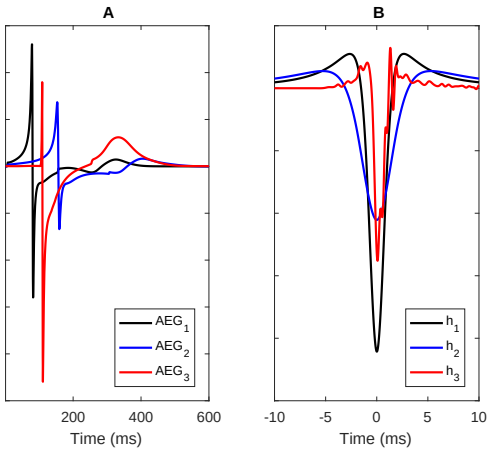


Figure 1: Three simulated atrial electrograms in (A) with their corresponding transfer functions in (B). The symmetry in h_1 and h_2 is the indicator of isotropic and homogeneous tissue activated by a planar wavefront. However the lower amplitudes and the spread of h_2 in time indicate lower conductivity in the tissue. On the other hand, the asymmetry of extrema in h_3 is an indicator of anisotropic tissue and a curved depolarization wavefront. Deriving these conclusions directly from the electrograms seems much more complicated. For a better comparison of the transfer functions, they have been shifted to 0 with respect to the the activation time of the cell under the electrode.

Video Quality Assessment in Video Streaming Services: Encoder Performance Comparison

Rufat Alizada
University of Twente
Dept. EEMCS, Group TE
Drienerlolaan 5, 7522 NB Enschede, The Netherlands
r.alizada@student.utwente.nl

Abstract

Video streaming services over networks have increased significantly in the past decade. Maximizing end viewers quality of experience (QoE) became more crucial requirement than the quality of service (QoS) awareness when deploying new broadcast platforms for the provisioning of high-quality streaming services. Since the majority of current multi-user video streaming services are encoded and transmitted through bandwidth-limited networks, proper encoder settings are required to satisfy the total channel rate constraint. In this work a QoE-driven High Efficiency Video Coding (HEVC) encoder adaptation scheme is proposed, aiming to measure overall acceptability of video content as perceived subjectively and maximize QoE of all the users for broadcasting networks. First, the influence of HEVC encoder on video streaming is investigated. The encoding and compression efficiency of the new standard in comparison to its predecessor H.264/AVC is given. Afterwards, a QoE-maximized encoder adaptation framework is formulated based on the obtained encoder parameter model. It turned out that the proper deployment of such framework on HFC (hybrid fiber-coaxial) network may result in average bit-rate reduction of 44% and average cost savings of 20%-60%.

1 Introduction

Video Streaming has become one of the most popular applications over next-generation networks. Interest in multimedia content and services on demand has been increased significantly, creating the need for the high quality content provision and effective compression techniques. Well accustomed to a variety of multimedia devices, consumers want a flexible digital lifestyle in which high-quality multimedia content follows them wherever they go and on whatever device they use. To meet this industry requirement in a way that interoperability is reassured, standardization activities have been taken place for the various video encoding techniques. Along with the rapid development of video compression standards and network transmission technologies, video streaming application has faced to a situation where the end-user expects a high quality of experience (QoE) available. QoE, defined by ITU-T [1] as a measure of the overall acceptability of an application or service, as perceived subjectively by the end-user, is the ultimate measure of user satisfaction to be maximized in a multimedia application. Due to the bandwidth-limited nature of wireless channels, it is essential to develop efficient video compression and transmission schemes to maximize QoE of real-time video streaming applications.

High efficiency video coding (HEVC) [2], also known as H.265/MPEG-H Part 2 is being rapidly adopted and is the last generation of video coding standard finalized in January 2013 by The Joint Collaborative Team on Video Coding (JCT-VC). The HEVC is developed within the constraints imposed by real-time processing and also,

will further reduce by 50% the bit rate required for high-quality video encoding compared to the previous related work H.264/AVC standard [2]. This increase is achieved due to new and improved tools implemented in the HEVC standard, such as the highly flexible and efficient block partitioning structure, larger prediction blocks, and precise inter/intra predictions. The new block named the in-loop sample adaptive offset (SAO) filter and the algorithm of entropy encoder called Context Adaptive Binary Arithmetic Coder (CABAC).

In this paper, a QoE-driven HEVC encoder adaptation framework is formulated based on the obtained encoder parameters and objective QoE model. Within proposed framework, the assessment of the HEVC encoders implemented on hybrid fiber coaxial is performed in terms of video quality and coding performance. For this reason, a set of video signals is used as input to the reference encoders. The content of the experimental set covers different spatiotemporal activity levels, making the assessment framework of this paper to examine performance, especially in the cases where two different codecs are benchmarked.

The paper organized as follows: Section 2 presents an overview of the HEVC encoder. Section 3 contains the description of the test methodology and assessment setup. Then, the detailed experimental results are presented in Section 4, and this paper is concluded in Section 6.

2 Overview of HEVC encoder

Today, H.264/MPEG-4 AVC is the dominant video coding technology used worldwide. As a rough estimate, about half the bits sent on communication networks worldwide are for coded video using AVC, and the percentage is still growing [2]. However, the emerging use of HEVC is likely to be the inflection point that will soon cause that growth to cease as the next generation rises toward dominance.

Due to the popularity of HD video and the growing interest in Ultra HD (UHD) formats [3] with resolutions of, for example, 3840x2160 or even 7680x4320 luma samples, HEVC [4] has been designed with a focus on high-resolution video. Even though the coding of HD and UHD video was one important aspect in the HEVC development, the standard has been designed to provide an improved coding efficiency relative to its predecessor AVC for all existing video coding applications.

HEVC standard is designed along the successful principle of block-based hybrid video coding. Following this principle, a picture is first partitioned into blocks and then each block is predicted by using either intra-picture or inter-picture prediction. While the former prediction method uses only decoded samples within the same picture as a reference, the latter uses displaced blocks of already decoded pictures as a reference. Since inter-picture prediction typically compensates for the motion of real-world objects between pictures of a video sequence, it is also referred to as motion-compensated prediction. While intra-picture prediction exploits the spatial redundancy between neighboring blocks inside a picture, motion-compensated prediction utilizes a large amount of temporal redundancy between pictures. In either case, the resulting prediction error, which is formed by taking the difference between the original block and its prediction, is transmitted using transform coding, which exploits the spatial redundancy inside a block and consists of a decorrelating linear transform, scalar quantization of the transform coefficients and entropy coding of the resulting transform coefficient levels.

Each picture in HEVC is subdivided into disjunct square blocks of the same size, each of which serves as the root of a first block partitioning quadtree structure, the coding tree, and which are therefore referred to as coding tree blocks (CTBs). The CTBs can be further subdivided along the coding tree structure into coding blocks (CBs), which are the entities for which an encoder has to decide between intra-picture and motion-compensated prediction. While increasing the size of the largest supported

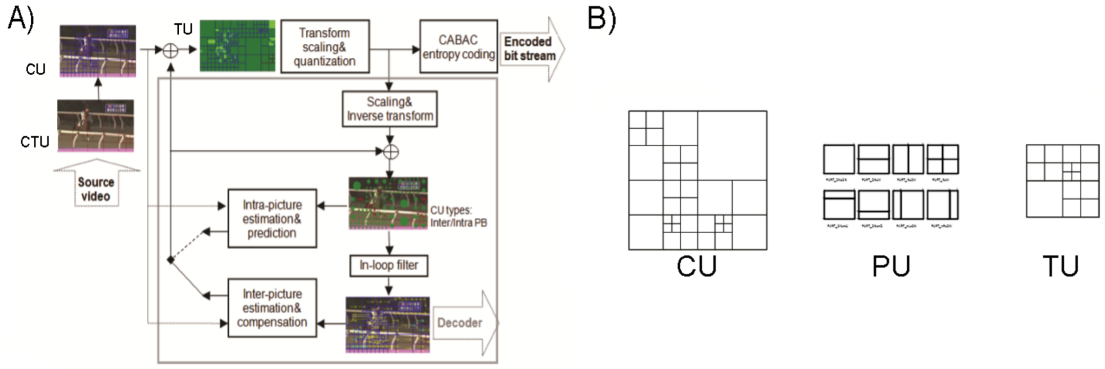


Figure 1: Picture partitioning example of coding quadtree CTU into CU (CB), partition modes for PU (PB), and transform quadtree within CU (TB).

block size is advantageous for high-resolution video, it may have a negative impact on coding efficiency for low-resolution video, in particular if low-complexity encoder implementations are used that are not capable of evaluating all supported sub-partitioning modes. For this reason, HEVC includes a flexible mechanism for partitioning video pictures into basic processing units of variable sizes.

A schematic description of the whole HEVC encoder block diagram is given in Figure 1 [5]. It receives generally an input YUV frame type and generates encoded bitstream data on its output. The HEVC encoder is based on different coding tools with high computational complexity compared to its previous standard.

2.1 Block-Based Coding

The HEVC continues to implement the block-based hybrid video coding framework, with the exception of the increased macroblock size (up to 64x64) compared to AVC. Three novel block concepts are introduced, namely: the Coding Unit (CU), the Prediction Unit (PU) and the Transform Unit (TU). CU is the basic coding unit similar to the H.264/AVCs macroblock and can have various sizes but is restricted to be square shaped. PU is the basic unit for prediction, where the largest allowed PU size is equal to the CU size. Other allowed PU sizes depend on prediction type, where asymmetric splitting options for inter-prediction is also considered. Finally, TU is the basic unit for transform and quantization, which may exceed the size of PU, but not that of the CU.

The general outline of the coding structure is formed by various sizes of CUs, PUs, and TUs in a recursive manner, once the size of the Largest Coding Unit (LCU) and the hierarchical depth of CU are defined. Given the size and the hierarchical depth of LCU, CU can be expressed as a recursive quadtree representation as it is depicted in Figure 3B, where the leaf nodes of CUs can be further split into PUs or TUs.

2.1.1 Intra-Prediction in HEVC

Intra prediction in HEVC is designed to efficiently model different directional structures typically present in video and image content. The set of available prediction directions has been selected to provide a good trade-off between encoding complexity and coding efficiency for typical video material. The sample prediction process itself is designed to have low computational requirements and to be consistent across different block sizes and prediction directions. This has been found especially important as the number of block sizes and prediction directions supported by HEVC intra coding far exceeds

those of previous video codecs, such as H.264/AVC. In H.264 standard, nine modes of prediction exist in a 4x4 block for intra prediction within a given frame and nine modes of prediction exist at the 8x8 level. It is even fewer at the 16x16 block level, dropping down to only four modes of prediction. Intra prediction attempts to estimate the state of adjacent blocks in a direction that minimizes the error of the estimate. In HEVC, a similar technique exists but the number of possible modes is 35-in line with the additional complexity of the codec (Figure 2). This creates a dramatically higher number of decision points involved in the analysis, as there are nearly two times the number of spatial intra-prediction sizes in HEVC as compared to H.264 and nearly four times the number of spatial intra-prediction directions.

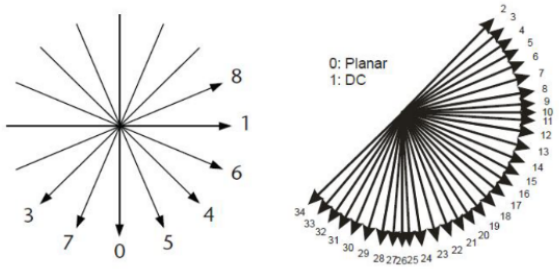


Figure 2: AVC vs HEVC Intra Prediction Modes.

2.1.2 Inter-Prediction in HEVC

The inter prediction in HEVC uses the frames stored in a reference frame buffer (with a display order independent prediction, as in AVC), which allows multiple bidirectional frame reference. A reference picture index and a motion vector displacement are needed in order to select reference area. The merging of adjacent PUs is possible, by the motion vector, not necessarily of rectangular shape as their parent CUs. In order to achieve encoding efficiency, skip and direct modes similar to the AVC ones are defined, and motion vector derivation or a new scheme named motion vector competition is performed on adjacent PUs. Motion compensation is performed with a quarter-sample motion vector precision. At TU level, an integer spatial transform (with the range from 4x4 to 64x64) is used, similar in concept to the DCT transform. In addition, a rotational transform can be used for block sizes larger than 8x8, and apply only to lower frequency components.

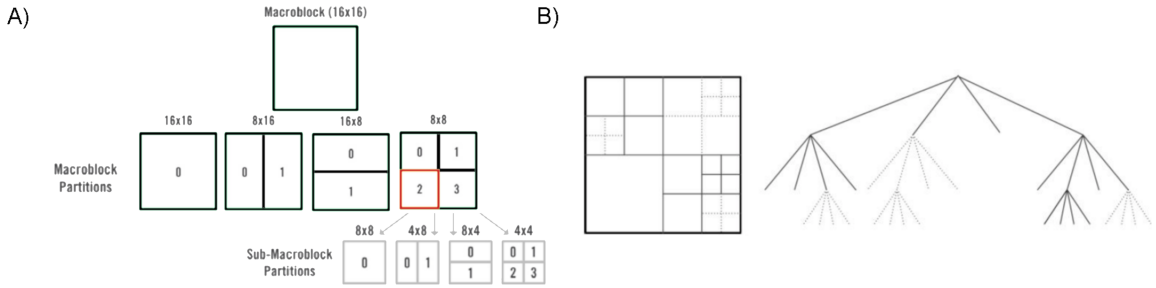


Figure 3: Inter Prediction and Coding: a) AVC Macroblock Partitions for inter prediction b) HEVC Quadtree Coding Structure for inter prediction.

2.1.3 Entropy Prediction in HEVC

Entropy coding is a lossless compression scheme performed at the last stage of video encoding (and first stage of video decoding), after the video signal has been reduced to a series of syntax elements. Syntax elements describe how the video signal can be reconstructed at the decoder. This includes the method of prediction (e.g., spatial or temporal prediction) along with its associated prediction parameters as well as the prediction error signal, also referred to as the residual signal. These syntax elements describe properties of the the aforementioned CU, PU and TU and loop filter (LF) of a coded block of pixels. The LF syntax elements are sent once per largest coding unit (LCU), and describe the type (edge or band) and offset for sample adaptive offset in-loop filtering.

Context-Based Adaptive Binary Arithmetic Coding (CABAC) [6] is a form of entropy coding used in H.264/AVC [7] and also in HEVC [4]. In H.264/AVC, CABAC provides a 9% to 14% improvement over the Huffman-based CAVLC [8]. CABAC involves three main functions: binarization, context modeling, and arithmetic coding. Binarization maps the syntax elements to binary symbols (bins). Context modeling estimates the probability of the bins. Finally, arithmetic coding compresses the bins to bits based on the estimated probability.

3 Test Methodology

This paper assesses the efficiency of recent implementations of the video encoders along two dimensions: the video quality obtained when a video signal is decoded at the receiver, and the computational complexity of the encoding and decoding processes. For performing the detailed performance analysis and in order to be as fair as possible due to the significant difference in the capabilities of the individual encoders, very similar settings for all tested encoders were used.

Below, the test methodology and the evaluation setup are explained in detail. Particularly, in Sub-Section 3.1, the VQ assessment is discussed, followed by the discussion of compression efficiency assessment, in Sub-Section 3.2. Finally, Sub-Section 3.3 gives an overview of the testing platform performed in HFC network.

3.1 Video Quality Assessment

In this section, firstly, the most popular current test methodologies for video quality assessment are reviewed and classified. Currently, there are many image and video quality assessment methods, each meeting different purposes. These methods can be classified in different ways depending on the criteria set adopted, as illustrated in Figure 4.

Currently, QoE under the subject of video quality assessments is roughly divided into two section that none other than Objective and Subjective methods. Most used subjective video quality assessment methods are described in the ITU recommendations ITU-R BT-500 [9] and ITU-T P.910 [10]. Aforementioned methods are focused on multimedia services. These tests are generally conducted under laboratory conditions, in which the supervisor explains the test instructions to the assessors. Later, assessors watch a test video; they grant an adjective score using 5-point MOS scale described in the Absolute Category Rating (ACR) method standardized in ITU-T Recommendation P.910. Although running subjective tests for video quality evaluation is the main way of evaluating VQ, it is rather an expensive, time consuming and tedious procedure. All this makes applying objective video metrics the best option for evaluating VQ for real-time applications.

Objective video quality assessment methods can be classified by using several considerations. Depending on the type of application service, objective methods are di-

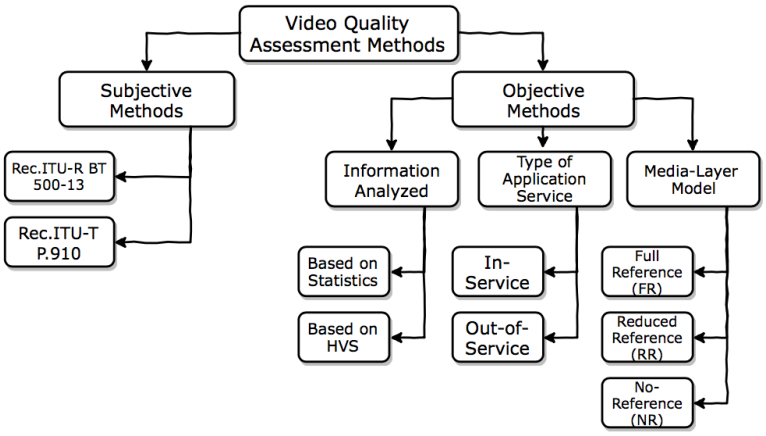


Figure 4: Classification of video quality assessment methods using different criterions.

vided into two categories [11]: (1) In-service methods: Real time VQ assessment applications with time constraints [12], [13]. (2) Out-of-service methods: Do not have time restrictions and are used in different tasks, such as video codec performance evaluation and video streaming services [14], [15]. In addition, as illustrated in Figure 5, the media-layer objective quality assessment methods can be further categorized as full-reference, reduced-reference, and no-reference [13] depending on whether a reference, partial information about a reference, or no reference is used in assessing the quality, respectively.

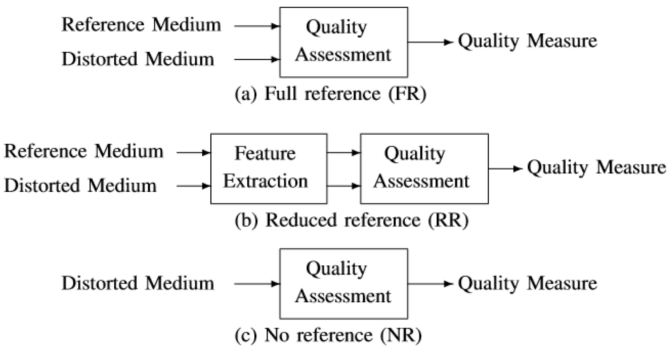


Figure 5: Overview of media layer models.

As illustrated in Figure 4, the category of information analyzed is sub-classified according to the information analyzed. The well known objective metrics, Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) are calculated based on statistical analysis of the pixels information. In turn, the metrics Structural Similarity (SSIM) [16], Video Quality Metric (VQM) [17] and algorithms based on Region of Interest (RoI) [18] or attentions maps [19], [20] are based on the Human Visual System. The overview of the objective video quality metrics is included in the Section 3.3.

For the evaluation of the VQ, four reference video clips were chosen out of recently designed LIVE-Netflix Video QoE Database, with content that represents various levels of spatial and temporal activity. A representative snapshot of each signal is depicted on Figure 6. The test signals have spatial resolutions of 1920x1080 (HD), 3840x2160 (QFHD) and 4096x2160 (UHD) and frame rates of 30 to 60 fps. For the experimental need of this paper, test signals were encoded from their original uncompressed YUV

format to AVC and to HEVC profiles. In order to maintain and achieve ideal comparison between the various profiles, it is necessary all the profile configurations have identical or very similar parameter values.

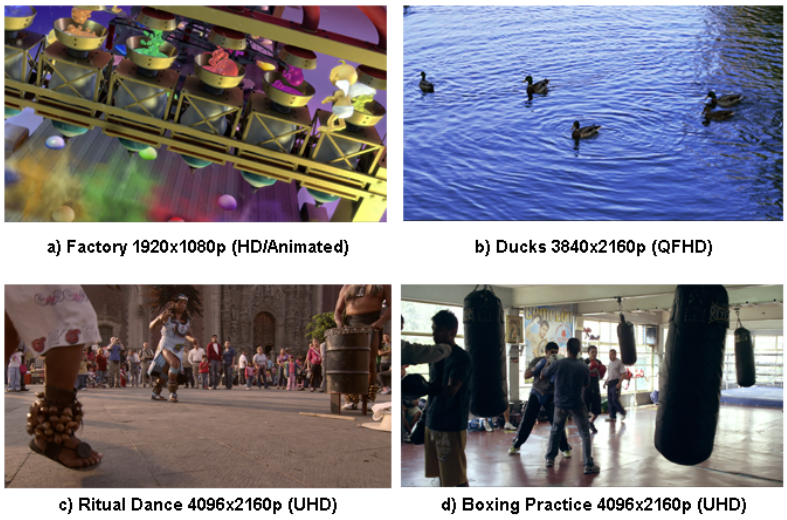


Figure 6: Representative frames taken from the test signals.

Main objective VQ metric used in this research is a reduced-reference video QoE measure, SSIMPLUS, that provides real-time prediction of the perceptual quality of a video based on human visual system behaviors, video content characteristics (such as spatial and temporal complexity, and video resolution), display device properties (such as screen size, resolution, and brightness), and viewing conditions (such as viewing distance and angle) [21]. SSIMPLUS assessment model combines the most significant HVS features, which include spatial frequency sensitivity, luminance masking, texture masking, temporal frequency sensitivity, and short-term memory effect. Moreover, it has a simple and clear structure and is easy for software implementation.

3.2 Compression Assessment

Compression efficiency is the most fundamental driving force behind the adoption of modern digital video compression technology, and HEVC is exceptionally strong in that area. However, it is also important to remember that the standard only provides encoders with the ability to compress video efficiently, it does not guarantee any particular level of quality since it does not govern whether or not encoders will take full advantage of the capability of the syntax design. As it was already mentioned, AVC and HEVC codecs follow the "block-based hybrid" coding approach. This type of coding exploits the spatial and temporal redundancy of the video frames. Frames are divided into three types: (1) I-frames (Intra Coded Picture) serve as an anchor for other frames to be decoded; (2) P-frames (Predictive Coded Picture) are predicted in a temporal manner only from previous frames (P-or I-frames); (3) B-frames (Bidirectional Coded Picture) are predicted from predicted from previous and following frames (I- P- or even B-frames) and achieve highest compression rate. The frames contained between two consecutive I-frames are called Group of Picture (GOP) [22]. The visual quality usually decreases with the GOP size. Thus, investigating the effect of encoder settings, such as input sequence resolution, frame rate, the length of GOP and quantization parameter (QP), on HEVC encoded video quality is necessary and challenging.

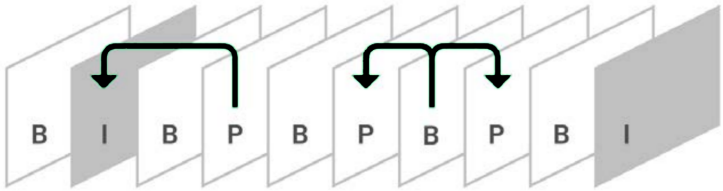


Figure 7: A typical sequence with I-, B- and P-Frames.

For the purpose of this analysis, multiple video sequences were selected and evaluated. All videos contain a non-static picture in a duration of ten seconds. One of the test sequences, named *Ducks* contains scenes of multiple ducks taking off from the surface of the water. This sample contains a large amount of changes and effects to stress various aspects of processing and test the stability of the encoder. Next two samples named *Ritual Dance* and *Boxing Practice* contain fast action artifacts and with the introduction of different amount scene changes and effects. The reason for these three different video sequences being evaluated is to determine if scene structure has some influence on the compression efficiency evaluation. In order to achieve accurate comparison, the GOP structure for all the encoding profiles and between two encoding methods consisted of the same I, P and B sequence, ensuring accurate benchmarking of both Intra and Inter-coding methods of AVC and HEVC profiles. The assessment of compression efficiency of HEVC is realized through objective measurement software Elecard StreamEye Tool by Elecard [23]. StreamEye is practical objective visual distortion measurement model for digital video compression. The primary purpose of StreamEye is to evaluate the video coding algorithms for the compression and visual quality through the comparison with the reference raw data.

3.3 Specification of testing platform

To evaluate how encoders perform, number of test sequences are prepared and the objective quality scores are collected. In contrast with previous studies, the sequences under test included ultra high definition footage which had been transcoded and compressed to varying degrees. The efficiency of the encoders is examined to extend to which the various objective metrics are compared and assessed. This section describes sequence preparation, gives an overview of used objective quality metrics and provides specifications of testing platform.

Table 1: Characteristics of test material used in the evaluation

File Name	Video Codec	Video Profile	Resolutions	fps	Bit Depth
Factory	AVC/HEVC	High@L5.1	1920x1080	30	8
Ducks	AVC/HEVC	High@L5.1	3840x2160	50	8
Ritual Dance	HEVC	Main@L5.1@Main	4096x2160	60	8
Boxing Practice	HEVC	Main@L5.1@Main	4096x2160	60	8

Thirty test video sequences are generated by subjecting four different original undistorted HD and UHD video sequences (*Factory*, *Ducks*, *Ritual Dance* and *Boxing Practice*) to AVC/HEVC coding schemes with different bitrates (2 to 5 Mbps for HD and 10 to 30 Mbps for UHD content). Selected bit rates are chosen to represent different real-life HDTV consumer and broadcasting applications from IPTV at the lower end to UHD on the upper end of the bitrate scale. Some target bit rates were rather aggressive

in order to be able to evaluate encoders at a point where they were stressed to process the content.

Evaluation of all test scenarios include five common objective video quality metrics. Chosen metrics are described below. Those metrics were chosen to represent a number of different approaches to the quality assessment problem. Aforementioned statistical metrics like PSNR and SSIM are included to provide a baseline against which the other metrics can be compared.

1. *Peak Signal-to-Noise Ratio (PSNR):* Traditional image quality assessment method. It is the ratio between the maximum power of the signal and the power of the difference signal between the reference and test images.

2. *Structural Similarity (SSIM)[16]:* Measures the structural similarity between the reference and test images based on the assumption that HVS is adapted for extracting structural information from a scene.

3. *Visual Information Fidelity (VIF)[25]:* Approaches quality evaluation by attempting to quantify distortion-induced differences in source information which can be usefully processed by the HVS. Scores range from 0 (worst) to 1 (best).

4. *Video Multi-Method Assessment Fusion (vmaf)[25]:* Objective FR VQ metric developed by Netflix. The metric is fusion of VIF and Detail Loss Metric (DLM), an image quality metric based on rationale of separately measuring the loss of details which affects the content visibility, and the redundant impairments which distracts viewer attention.

5. *Perceptual Fidelity (PF)[21]:* Novel SSIMPLUS QoE measure that provides straightforward predictions on what an average consumer says about the quality of the video content being delivered on a scale of 0-100 and also categories the quality as either bad, poor, fair, good or excellent.

First test scenario considers the direct output of the encoders. The compression efficiency evaluation of all test sequences is performed over the first test scenario. Second test scenario considers a video streaming service over hybrid fiber-coaxial (HFC) access network. In HFC network, the content is sent from the cable system's distribution facility to local communities through optical fiber subscriber lines. The tests were performed over this network in order to replicate real-life television broadcasting and simulate various cases of network conditions, that may cause "damage" of video stream, in particular, due to delay and packet losses in the network. Third test scenario considers the evaluation of the content by utilizing the device-adaptation feature of SSIMPLUS software [26]. This feature takes into consideration the fact that human quality assessment of the same video content can be significantly different when it is displayed on different viewing devices, such as HDTV, digital TV, projectors, PCs, and smartphones, and many more. The outcome of the third test scenario could be used to adapt video QoE analysis to any display device and viewing conditions.

In case of the first scenario, the reference software was used for both AVC and HEVC coding. For this work, the FFmpeg library libx264 which is the x264 H.264/MPEG-4 AVC encoder wrapper and libx265 which is the x265 H.265/HEVC encoder wrapper were used to encode the content. FFmpeg is an easy to use open source software capable of performing a wide range of multimedia operations including transcoding, encoding and conversion of audio/video content [27]. In this work, FFmpeg version 3.4.1, built with Apple LLVM version 9.0 was used for encoding the videos. Encoding was performed on a MacBook Pro laptop with 8 GB RAM, Intel Iris and Intel Dual Core i5@2.70GHz running 64-bit macOS High Sierra v.10.1.3. Rest of the evaluation was carried out on a AWS Elemental Server rack, also equipped with Intel multi-core technology. This Intel platform is composed of two Intel Core i7-6820HQ processors,

running at 2.7 GHz, and 16GB of DDR4 memory. AWS Elemental Live software v.4.0 is implemented over the server and provides support for all recent audio/video coding schemes.

4 Evaluation

4.1 Video Quality of HEVC

For obtaining experimental results, most of the test sequences were selected according to the encoder test conditions, as presented in Table 1. All sequences include different texture and motion characteristics to bring up a reasonable relationship between bit rate, PSNR, and encoding settings.

Figure 8 illustrates VQ curves of HEVC and x264 encoders for two typical examples of tested sequences. As it is clearly seen from Figure 8, the HEVC encoders provides significant gains in terms of perceived quality compared to its predecessor AVC.

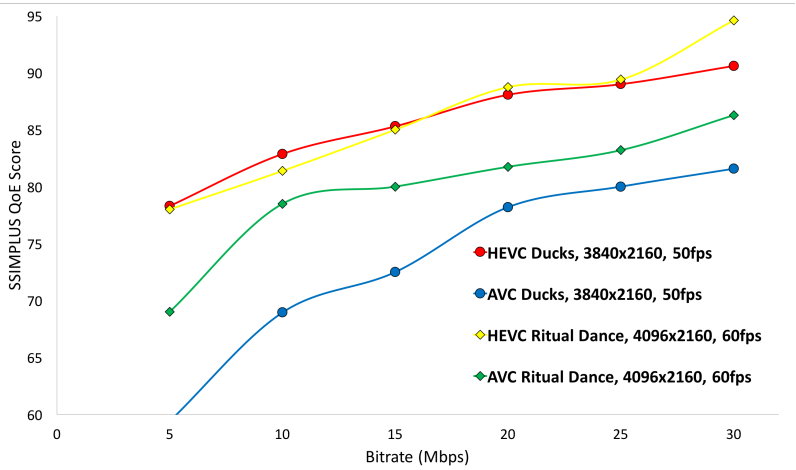


Figure 8: Video quality assessment for several typical examples of tested sequences.

Table 2 illustrates the results comparing all possible conditions for the 1080p and 4k content, respectively. Comparing HEVC and AVC at similar bit rates, HEVC always provides statistically better visual quality when compared to AVC for *Ducks*, *Ritual Dance* and *Boxing Practice*. The table also includes "Motion" metric, a simple measure of the temporal difference between adjacent frames. The score typically ranges from 0 (static) to 20 (high-motion). This simple feature illustrates that content under assessment contains a large amount of changes and effects to stress various aspects of the processing of encoder. For the animated content *Factory*, there is not sufficient statistical evidence to show that HEVC outperforms AVC, especially at high bit rates.

4.1.1 QoE driven bandwidth optimization

Once the quality of content is evaluated, many benefits come as a natural next step. One of such benefits is bandwidth optimization. Although a significant number of solutions have been proposed in the industry for saving bandwidth, talking about bandwidth reductions without maintaining the right level of visual QoE makes little sense. Due to the lack of proper QoE assessment tools, existing bandwidth saving approaches, whether it is applied to encoding/transcoding or streaming optimization, result in unstable results. To perform bandwidth optimization properly, the first step

Table 2: Video quality assessment of 1080p and UHD content encoded at 10, 20 and 30 Mbps. Including the comparison of the HEVC and AVC encoding schemes.

		HEVC			AVC		
Content	**VQM**	10Mpbs	20Mpbs	30Mpbs	10Mpbs	20Mpbs	30Mpbs
QFHD 3840x2160 Ducks 500 frames Motion: 4.84	PSNR	27.7	28.8	29.3	25.59	26.79	27.4
	SSIM	0.67	0.71	0.73	0.59	0.64	0.66
	vmaf	48.36	58.57	64.42	33.69	42.25	46.8
	PF	82.87	88.07	90.6	68.95	78.2	81.57
	VIF	0.68	0.72	0.74	0.61	0.64	0.67
UHD 4096x2160 Ritual Dance 600 frames Motion: 16.59	PSNR	27.4	28.7	29	26.1	26.4	27.2
	SSIM	0.64	0.69	0.71	0.63	0.65	0.67
	vmaf	46.45	59.66	62.51	43.11	47.2	54.68
	PF	81.38	88.74	94.61	78.49	81.74	86.27
	VIF	0.67	0.72	0.8	0.64	0.67	0.71
UHD 4096x2160 Boxing Practise 600 frames Motion:12.13	PSNR	27.4	28.3	28.9	27.9	28	28.1
	SSIM	0.67	0.7	0.72	0.64	0.66	0.69
	vmaf	48.39	58.98	61.41	41.08	48.53	55.42
	PF	84.78	87.26	92.95	74.81	84.06	87.43
	VIF	0.69	0.7	0.76	0.62	0.68	0.7

has to be adopting a trusted QoE metric with powerful functionalities, e.g., accurate, meaningful and consistent quality assessment cross resolutions, frame rates, dynamic ranges, viewing device and video content. An illustrative example using SSIMPLUS as the example QoE metric is given to demonstrate how large bandwidth savings can be achieved in live and file-based operations by making use of such a QoE metric.

Significant bandwidth savings can be obtained by adopting a QoE measure that produces consistent QoE assessment across content, resolution, and user device, each of which could lead to significant gain. Firstly, because of the difference in encoding difficulty of different content (Sample 3 and Sample 4 shown in the Fig.9), to reach a guaranteed QoE quality level (SSIMPLUS = 90), using a fixed bandwidth to encode all videos may be a waste, depending on video content, e.g., using a fixed 30 Mbits when only 27 Mbits is necessary for Sample 4.

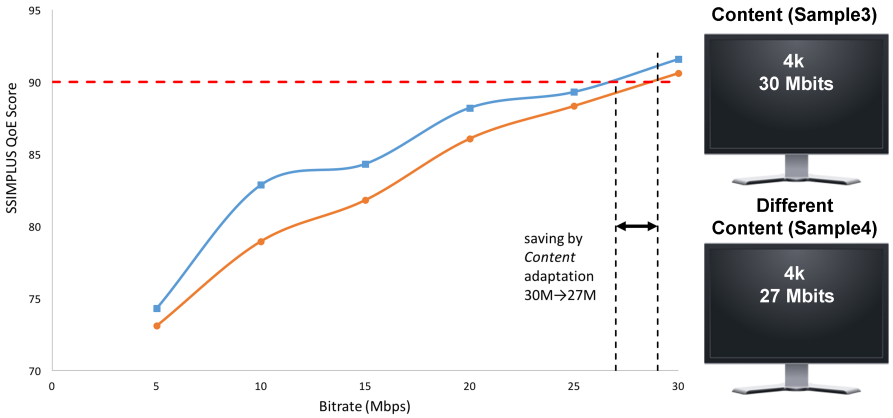


Figure 9: Illustration of how bandwidth savings is achieved by using a QoE metric that is able to adapt to video content.

Second, when the same content is encoded to two or more spatial/temporal resolutions, the capability of picking the most cost-effective spatial/temporal resolution to achieve the guaranteed quality level can also help save large bandwidth, e.g., a bandwidth reduction from 3.4 Mbits to 2.3 Mbits is obtained by switching from 1080p to 720p resolutions, as shown in the Figure 10.

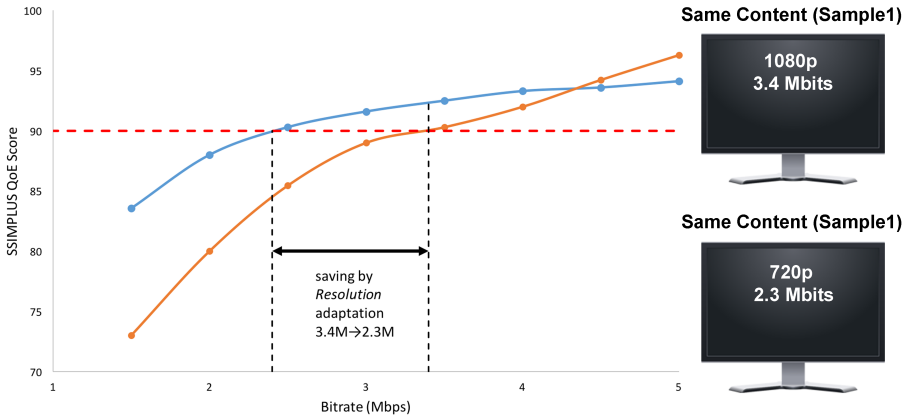


Figure 10: Illustration of how bandwidth savings is achieved by using a QoE metric that is able to adapt to video resolution.

Finally, the perceptual QoE varies significantly on different viewing devices. This is illustrated in the quality-bitrate curve in the Figure 11, which shows that when the user is known to use a smartphone rather than a TV to watch the video, a bandwidth of 17 Mbits is sufficient to achieve the same target quality level (SSIMPLUS = 90). With all three factors combined, a total of 44% bandwidth savings may be obtained (from 30 Mbits to 17 Mbits).

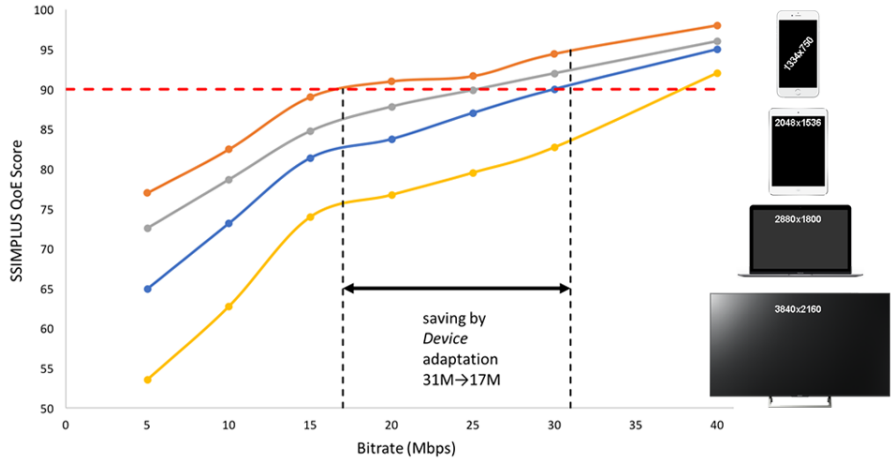


Figure 11: Illustration of how bandwidth savings is achieved by using a QoE metric that is able to adapt to user viewing device.

Although given examples are here for illustration purposes only, and in practice users may be constrained to explore all three factors for maximum cost-savings, conducted research suggests that for most video content and the most common usage profiles, an average cost saving of 20%-60% is typically achieved by properly adopting this QoE metric-driven bandwidth optimization technology. Such bandwidth savings can be implemented by adaptive operation of video encoders/transcoders, and may also be incorporated into adaptive streaming frameworks to achieve similar goals in a dynamic way.

5 Coding Efficiency of HEVC

For illustration, Figure 12 shows the partitioning of a picture with 3860x2160 luma samples into 16x16 macroblocks and 64x64 CTUs. It can be seen that 16x16 macroblock covers only a very small area of a picture, much smaller than the regions that can typically be described by the same motion parameters. Taking into account that some of the CTUs will be subdivided for assigning different prediction modes and parameters, the partitioning into 64x64 CTUs provides a more suitable description.

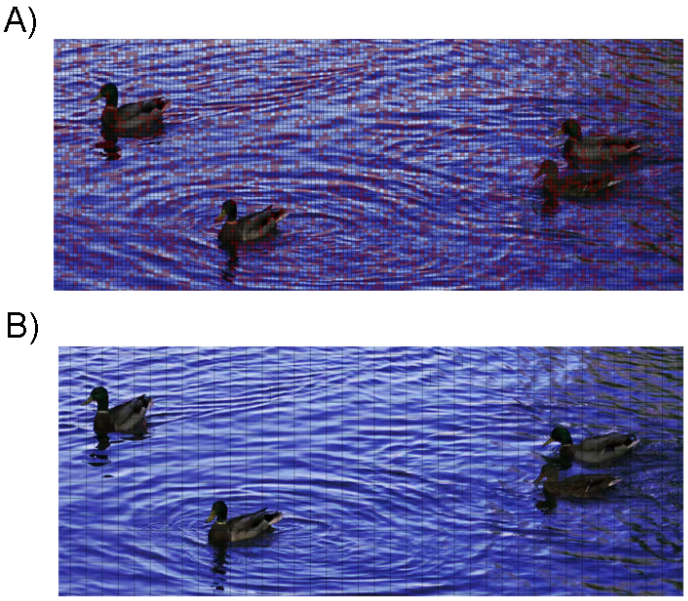


Figure 12: Illustration of the partitioning of a picture with 3840x2160 luma samples into macroblocks and coding tree units: (a) Partitioning of the picture into 16x16 macroblocks as found in all prior video coding standards of the ITU-T and ISO/IEC; (b) Partitioning of the picture into 64x64 coding tree units, the largest coding tree unit size supported in the Main profile of HEVC.

5.1 Performance Comparison of HEVC vs AVC

This section presents evaluation of the compression efficiency of the HEVC algorithm in comparison to the AVC for the test signals under test, when same encoding parameters have been selected.

The first step in compressing of content is to segregate the data into different classes. Depending on the importance of the data it contains, each class is allocated a portion of the total bit budget, such that the compressed data has the minimum possible distortion. This procedure is called bit allocation. Based upon the demanded bit rate and the current fullness of the buffer, a target bit rate for the entire GOP is determined, together with the QP for the GOP's I-picture and first P-picture. In video coding, it is expected that the encoder could adaptively select the encoding parameters to optimize the bit allocation to different sources under the given constraints. Table 3 illustrates the bit allocation of the encoders under the assessment. The experimental results for the x264 reference encoder illustrate that some frames in the assessed content were encoded using 20.2 Mbits, even tough target bitrate was set to 10 Mbits. In addition, Table 3 illustrates that AWS Elemental encoder has the most accurate rate controller that

tries to allocate available bit budget equally between video units (macroblock, frame, GOP). Proper usage of bit allocation data could be used for efficient management of available bit budget and accurate performance comparison of the encoders.

Table 3: Bit allocation of encoders under comparison for *Ducks* encoded at 10 Mbps

	Bit Allocation (Mbits)		
Encoders	Maximum	Average	Minimum
x264 AVC	20.2	10.5	9.2
x256 HEVC	11.9	10.1	8.7
AWS Elemental HEVC	10.4	10.1	10

To evaluate performance even further, rate-distortion assessment of HEVC in terms of the bit budget is performed. The bit rate reduction of one codec over another for a similar quality is estimated using the Bjøntegaard Delta Rate (BD-Rate) [28]. The Bjøntegaard model relies on PSNR measurements to determine the average bit-rate difference for the same objective quality. Since evaluation is performed with *yuv420* content, seperate PSNR values are obtained for luma (Y) and chroma (U,V) components. Combined $PSNR_{YUV}$ value are calculated as a weighted sum of the PSNR values per each frame of each individual component [29], as shown in Eq.1.

$$PSNR_{YUV} = \frac{6 \times PSNR_Y + PSNR_U + PSNR_V}{8} \tag{1}$$

Although a more realistic estimate of the performance efficiency can be obtained by considering subjective ratings instead of PSNR values. Using the combined $PSNR_{YUV}$ in rate-distortion assessment could be used to determine the trade-offs between luma and chroma component fidelity [29].

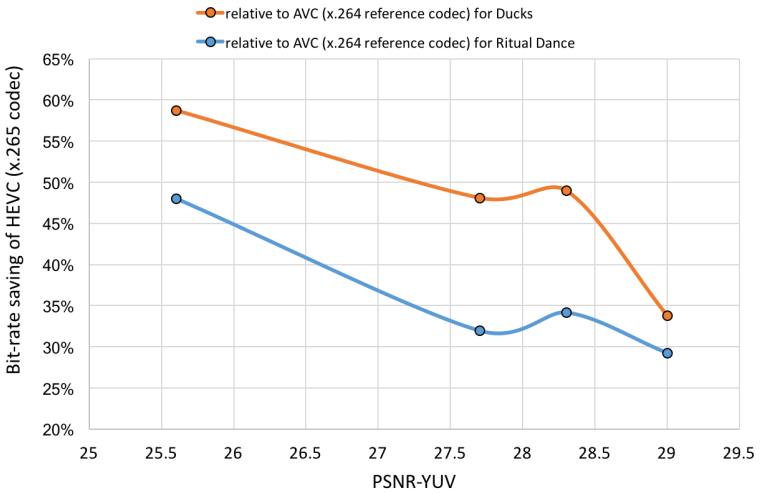


Figure 13: Bit-rate saving plots for several typical examples of tested sequences.

Figure 13 present rate-distortion curves of HEVC bitrate savings for two typical examples of tested sequences. As it is clearly seen, the HEVC provides significants gains in term of coding efficiency compared to H.264/AVC. Table 4 provides a summary of the bitrate reduction results, where negative BD-rate values indicate bitrate savings in contrast to positive values, which indicate the required overhead in bitrate to achieve the same $PSNR_{YUV}$ values.

Table 4: Summarized BD bit rate experimental results.

	BD-rate in %		
Encoders	x265 HEVC	x264 AVC	AWS Elemental HEVC
x256 HEVC		-47.3	19.2
x264 AVC	46.6		69.1
AWS Elemental HEVC	-21.4	-64.3	

As shown in Table 4, AWS Elemental HEVC outperforms both reference codecs. The average BD bit rate savings of AWS Elemental HEVC encoder relative to AVC and HEVC (x265) are 64.3% and 21.4%, respectively. As it is also observed from Table 4, the bit rate savings, on average, the HEVC (x265) encoder achieves an average gain of 47.3% in terms of bit-rate savings compared to AVC. Proper usage of BD bit rate values could be used for accurate performance comparison of the encoders.

5.2 Encoding Efficiency of HEVC

This section presents the experimental results of the comparison between HEVC and AVC encoded signals. Across the encoding process of both the HEVC and AVC profiles, all encoding parameters remained identical in order to quantitatively compare encoding efficiency of the encoder. Figure 14 illustrates the VQ evaluation per single frame, without focusing only on the average results. This facilitates the full-length content comparison together with opportunities to highlight the frames with higher complexity and discover anomalies in the transcoded files.

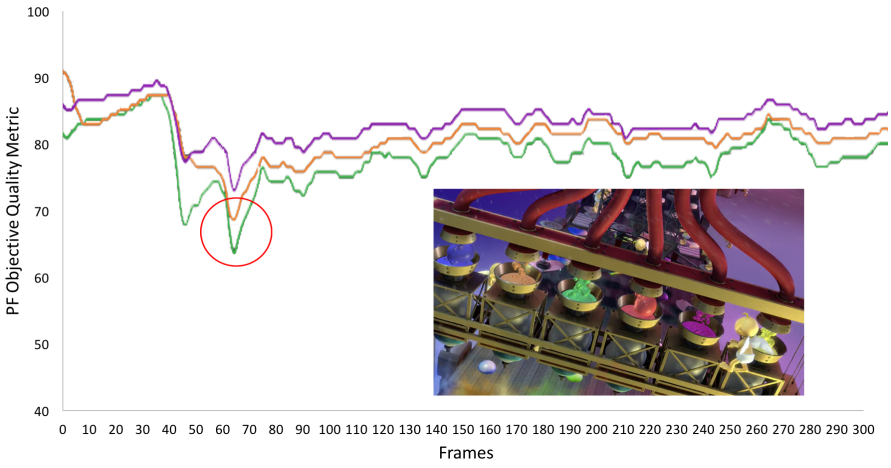


Figure 14: Per-frame VQA for several codecs.

Measurement values illustrated in Figure 14 are different, but the behaviour is similar between multiple codecs. This means that the most of the peaks are on the same scene/frame of the asset. The red circle highlights the scene or frame(s) with the highest complexity to transcode. The screenshot of this particular frame is also illustrated in Figure 14. This particular frame of animated content contains relatively fast moving objects with the complex color background, which explains why tested encoder struggles to meet its bit budget.

As it can be observed from experimental results, the encoding efficiency of the HEVC encoded signals appear to be better compared to the AVC/H.264. This is in

line with the objective of the HEVC to maintain the encoding efficiency compared to its predecessor AVC, while almost doubling compression and performance efficiency of the bitstream, as shown in the previous section. Commenting further on the experimental results, although the average PSNR score of the HEVC is similar to the respective one of the AVC in some cases, PF results illustrate that the HEVC encoding performance appears to have greater variance with higher QoE, that outperforms instantly the AVC performance.

6 Conclusions

This paper presents a detailed description of the objective quality evaluation tests conducted to benchmark the performance of HEVC and AVC video codecs for real-time video applications. The evaluation was performed using various HD and UHD content encoded at various bit rates. High accuracy software assessment tools were used to accurately compare the performance of the investigated codecs. Evaluation of test results shows that HEVC offers improvements in compression performance when compared to AVC, if one considers a wide range of bit rates from low to high, corresponding to video with low to transparent quality. More specifically, objective based QoE measurements show that average bit-rate reduction of 43% and average cost saving of 20%-60% is typically achieved by properly adopting this QoE metric-driven bandwidth optimization technology.

Acknowledgments

The work in this paper has been performed as a part of an internship at VodafoneZiggo, the Netherlands. The author thanks Wilhelm Zijlstra and whole Apps Engineering team of VodafoneZiggo, without which this research could not have been conducted. The author also wishes to thank Prof. Raymond Veldhuis for the patient guidance, encouragement and advice he has provided throughout the internship.

Finally, the author wishes to thank SSIMWAVE Inc.and ELECARD for their guidance and technical support as well as for providing necessary tools for this research.

References

[1] (ITU-T), *"ITU-T Recommendation G.1070 Opinion model for video telephony applications"*, Tech. Rep.,2012

[2] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 22, no. 12, pp. 16491668, 2012.

[3] ITU-R Rec. BT.2020 (2012) Parameter values for ultra-high definition television systems for production and international programme exchange.

[4] ITU-T Rec. H.265 and ISO/IEC 23008-10 (2013) High efficiency video coding.

[5] B. Bross, W.-J. Han, J.-R. Ohm, G. J. Sullivan, and T. Wiegand, "High efficiency video coding (HEVC) text specification draft 8," JCTVC-J1003 July, 2012.

[6] Marpe D, Schwarz H, Wiegand T, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard," IEEE Trans CSVT 13(7):620636 (2003)

[7] ITU-T Rec. H.264 and ISO/IEC 14496-10 (2003) Advanced video coding

[8] Alshina E, Alshin A, "Multi-parameter probability up-date for CABAC, Joint Collaborative Team on Video Coding (JCT-VC)", Document JCTVC-F254, Torino, July 2011

[9] ITU-R Recommendation BT.500, Methodology for the Subjective Assessment of the Quality of Television Pictures, ITU-T, Geneva, Switzerland, Jan. 2012.

[10] ITU-T Recommendation P.910, Subjective Video Quality Assessment Methods for Multimedia Applications," ITU-T, Geneva, Switzerland, Apr. 2008.

[11] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam, "Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison.," IEEE Trans. on Broadcasting, vol. 57, no.2, pp. 165-182, Jun. 2011.

[12] K. Yamagishi and T. Hayashi, "Parametric packet-layer model for monitoring video quality of IPTV services," in Proc. Int. Conference of Communications, pp. 110-114, Beijing, China, May 2008.

[13] M. Garcia and A. Raake, "Impairment-factor-based audio-visual quality model for IPTV," in Proc. International Workshop on Quality Multimedia Experience, pp. 1-6, California, US., Jul. 2009.

[14] M. Martines, M. Lopez, P. Pinol, M. Malumbres, and J. Oliver, "Study of Objective Quality Assessment Metrics for Video Codec Design and Evaluation," in Proc. IEEE International Symposium on Multimedia, pp. 517-524, California, US., Dec. 2006.

[15] B. Ciubotaru, G.-M. Muntean, and G. Ghinea, "Objective assessment of region of interest-aware adaptive multimedia streaming quality," IEEE Trans. on Broadcasting, vol. 55, no. 2, pp. 202-212, Jun. 2009.

[16] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Processing 13, 600612 (Apr. 2004).

[17] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," IEEE Trans. Broadcast., vol. 50, no. 3, pp. 312322, Sep. 2004.

[18] H. Kwon, H. Han, S. Lee, W. Choi, and B. Kang, "New Video Enhancement Preprocessor Using the Region-Of-Interest for the Videoconferencing," IEEE Trans. Consumer Electron., vol. 56, no. 4, pp. 2644-2651, Nov. 2010.

[19] J. You, A. Perkis, M. Gabbouj, and M. M. Hannuksela, "Perceptual quality assessment based on visual attention analysis," in Proc. International Conference on Multimedia, pp. 561564, Beijing, China, May 2009.

[20] A. K. Noorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," IEEE J. Select. Topics Signal Processing, vol. 3, no. 2, pp. 193201, Apr. 2009.

[21] SSIMPLUS: The most accurate video quality measure, https://www.ssimwave.com/from-the-experts/ssimplus-the-most-accurate-video-quality-measure/

[22] Sze V, Budagavi M, Sullivan G.J., "High Efficiency Video Coding (HEVC), Algorithms and Architectures," Springer; 2016.

[23] ELECARD StreamEye: Video analysis test software, https://www.elecard.com/products/video-analysis/streameye

[24] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Temporal effects on subjective video quality of experience," Transactions on Image Processing, under review.

[25] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric." http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html

[26] A. Rehman, K. Zeng and Z. Wang, "Display device-adapted video quality-of-experience assessment," IST/SPIE Electronic Imaging: Human Vision Electronic Imaging, Feb.2015.

[27] FFMPEG. FFmpeg and H.264 Encoding Guide. https://trac.ffmpeg.org/wiki/Encode/H.264

[28] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves", ITU-T Q.6/SG16 VCEG 13th Meeting, Document VCEG-M33, Austin, USA, Apr. 2001.

[29] J. Ohm, G.J. Sullivan, H. Schwarz, T.K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standardsincluding High Efficiency Video Coding (HEVC)," Circuits and Systems for Video Technology, IEEE Transactions on , vol. 22, no.12, pp.1669-1684, Dec. 2012.

Operational Rate-Constrained Noise Reduction for Generalized Binaural Hearing Aid Setups

Jamal Amini[†], *Richard C. Hendriks*[†], *Richard Heusdens*[†], *Meng Guo*[⋆] *and Jesper Jensen*[⋆*]

[†]Circuits and Systems (CAS) Group, Delft University of Technology, the Netherlands
[⋆]Oticon A/S and Electronic Systems Department, Denmark
[*]Aalborg University, Denmark
[†]{j.amini, r.c.hendriks, r.heusdens}@tudelft.nl
[⋆]{megu,jesj}@oticon.com

Abstract

Using noise reduction algorithms, hearing aids (HAs) can increase the speech intelligibility for HA users. With a rapid growth in the use of wireless technology, HAs can now potentially be wirelessly linked to each other to build a binaural HA system. Multi-microphone noise reduction techniques can be used in such binaural HAs systems, which may further increase intelligibility in comparison with monaural noise reduction algorithms (K. Eneman 2008). Moreover, the binaural HA system can be generalized to a (small) wireless acoustic sensor network (WASN) by collaborating with other assistive (wireless) devices. Multi-microphone noise reduction (beamforming) can be performed in such a WASN by transmitting all microphone signals to a fusion center (FC), for instance to the left-side HA, and then estimating the sources of interest and suppressing the undesired sources (interferers).

As the transmission capacities of wireless links between devices are limited, the microphone signals first need to be compressed/quantized before being transmitted to the FC. Several optimal and sub-optimal approaches have been proposed to constrain the rate of transmission between the two HAs in the binaural HA setup. An optimal binaural rate-constrained beamforming algorithm is proposed in (O. Roy and M. Vetterli 2009). The method optimally trades off the rate against the mean square error (MSE) of Gaussian target signal estimation. However, the method assumes that there are only two processing nodes (two HAs), and thus transmission between the binaural HAs and other assistive devices is not taken into account. Moreover, the method is less applicable in practice as joint microphone signal statistics need to be known at all nodes and (infinitely) long-block continuous-range vector quantizers are assumed. To address the need for the knowledge of the joint statistics, sub-optimal methods have been proposed in (S. Srinivasan 2009) and (O. Roy and M. Vetterli 2009). However, these methods are inevitably acoustic scene dependent meaning that the performance may be severely affected by acoustic scene parameters such as target source location, spatial noise distributions, etc. Moreover, the performance does not (asymptotically) approach the optimal performance for high bit rates, and the loss in performance is significant, even at high bit rates.

From a more general (practical) perspective, the binaural rate-constrained beamforming problem can be generalized in two main aspects. First, unlike binaural approaches with two processing nodes, a (small) WASN can be considered and the rate-distortion tradeoff can be derived for the generalized setup. Second, existing beamforming algorithms (or any other processing strategy) can be utilized in a more sophisticated way, resolving the acoustic scene dependency issue of the sub-optimal approaches.

In this paper, we propose an operational rate-constrained noise reduction framework for optimal rate allocation and strategy selection across frequency. The proposed problem is based on the general setup of a (small) WASN. Unlike (O. Roy and M. Vetterli 2009), the joint statistics are assumed to be known only at the FC. Moreover, a discrete set $\mathcal{A} = \{A_1, A_2, \ldots, A_{N_A}\}$ of strategy candidates (could be different microphone selections, different algorithms, etc.) is designed to resolve the acoustic scene dependency issue of the existing sub-optimal approaches. The proposed method aims at optimally selecting the strategy candidate and distributing the total resource budget, say $R_{\max}$, to different frequency components in order to optimize a fidelity criterion (for example, minimizing the MSE between the target speech signal and its estimate). The proposed optimization problem is given by

$$\begin{aligned} \min_{\boldsymbol{\alpha}\in\mathcal{A}'} \quad & \min_{\mathbf{r}\in\mathcal{Q}} \quad D(\boldsymbol{\alpha},\mathbf{r}) \\ \text{subject to} \quad & R(\mathbf{r}) \leq R_{\max}, \end{aligned} \tag{1}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{N_f}]^{\mathrm{T}}$, with α_k a possible strategy choice for a particular frequency, and similarly for $\mathbf{r} = [r_1, \ldots, r_{N_f}]^{\mathrm{T}}$. $\mathcal{A}' = \{\boldsymbol{\alpha} \mid \alpha_k \in \mathcal{A},\ k = 1, \ldots, N_f\}$ denotes the set of all possible strategy choices. Similarly, $\mathcal{Q} = \{\mathbf{r} \mid p_k \leq r_k \leq q_k,\ k = 1, \ldots, N_f\}$ denotes the set of all possible rate allocations across frequency. The minimum and the maximum possible operating bit rates are denoted by p_k and q_k, respectively, for a particular frequency. The distortion function $D(\boldsymbol{\alpha}, \mathbf{r})$ is defined as the averaged power spectral density of the estimation errors in different frequencies, in terms of algorithm choice and rate allocation across frequencies. The (global constraint) function $R(\mathbf{r})$ is an average of all rates across frequency. The Lagrange multiplier (LM) technique from (H. Everett 1963) and (Y. Shoham 1988) is used for solving the optimization problem.

As an example, using uniform quantizers, both the optimal strategy choices and the optimal rate allocations are found for a specific generalized binaural HA setup (the binaural setup together with an assistive device). Based on the output MSE gap between the monaural (i.e., no communication) setup and the (rate-constrained) generalized binaural setup, the performance of the proposed method is evaluated. The results show that the proposed method outperforms significantly the methods with naive strategy selection and equal rate allocation across frequencies.

Breaking Out of the Black Box in Automated Flower Recognition

D.H. Apriyanti [1,2], L.J. Spreeuwers [1], R.N.J. Veldhuis [1]
[1]University of Twente
Data Management and Biometrics Group, Faculty of EEMCS
P.O. Box 217, 7500 AE Enschede, The Netherlands
[2]Indonesian Institute of Sciences (LIPI)
Purwodadi Botanic Garden
Jl. Raya Surabaya Malang Km. 65, Purwodadi, Pasuruan, Indonesia
diah007@lipi.go.id
{l.j.spreeuwers, r.n.j. veldhuis}@utwente.nl

Abstract

Currently, fully automated methods for plant recognition are being developed. These systems work based on images. But, they only give the name of the plants as a result and do not give an explanation (black box approach). On the contrary, giving the explanation (by describing the plants characteristics) formally is used by taxonomists before they can determine the plants name. Providing the explanation is very important since it fit in the standard practice and as a basis of plant identification for several decades.

This paper comes with the new perspective in automated image-based plant recognition. A system that acts not as a black box system has proposed. It can mimic the taxonomist in explaining the decision and giving useful alternatives. The research firstly focuses on the flower. In the first part of this paper, the background of the new perspective is presented. Then, some problems that we want to answer also will be shown. Finally, some solutions to achieve the goal are discussed. As the first attempts, we also conduct a little experiment using the decision tree to deal with the goal.

1 Introduction

In Indonesia, there is a need for automated plant recognition because the number of taxonomists is limited, while there is high biodiversity [1]. Not only in Indonesia, but also elsewhere in the world, botanists can benefit from an automated plant identification process. Plant identification provides useful input for the management of biodiversity, such as managing livestock systems, protecting threatened species from trading, understanding what will grow best in an area, etc.

To identify plants, taxonomists follow a systematic approach called the identification key. It is a list of characteristics that can lead taxonomists to the species name. They work according to this list, but because most identification keys are paper-based, the process requires access to literature, time, and skills [2, 3]. However, computer-based identification keys have already been developed, such as stand-alone applications like Delta Intkey, Lucid [4, 5] or interactive web applications like GoBotany, MEKA, FloraGator, NatureGate, etc [6, 7, 8, 9]. These systems are already easier to use but still require expert knowledge.

Nowadays, fully automated methods are also being developed [3]. These systems work based on images. They only give the name of the plants as a result and do not give an explanation (black box approach). Providing the explanation is important because

it can mimic the taxonomists work and fit in the standard practice. Actually, the explanation is not only useful for taxonomists/botanists, but also for non-specialists, for example for educational purposes. Another thing is about the decision. The decision provided by these systems are crisp. In fact, the decision come with uncertainty. Therefore, it will be good if there is a system that also provide alternatives. Both explanation and alternatives will improve the trustworthiness of the system.

Although there are some systems that give some explanations about the plant, but the explanation comes after the decision. They just picked the description of the plant from database after they know the decision. It is different with the real traditional plant identification.

So, the goal of this paper is to design an automated image-based system for plant identification that mimics the taxonomist in explaining the decision and giving useful alternatives. For that purpose, this paper will seek to answer some problems regarding to that goal, i.e. which types of architecture are useful/work, how to integrate taxonomist knowledge, and how to handle uncertainty?

To understand easily our proposed solution, we organized this paper into some sections. Section 1 is introduction (we have already been here). We explain the background and the goal of this research. Then, the remained of this paper are organized as follows. Section 2 describes the method that we proposed, with some sub sections to answer the problems. Section 3 contains a little experiment and result. Finally, the last section contains conclusion of this research.

2 Proposed Method

2.1 Types of possible architecture

To deal with the goal, we should make a bridge to accommodate the recent plant identification method in Computer Science and the method which taxonomists usually did in Taxonomy. For that purpose, we think that the most possible architecture to solve the problems decribed above are Decision Tree and Bayesian Network (BN). Both of them have the possibility of providing us with alternatives. They show steps that can explain the decision and also can handle uncertainty.

Decision tree is a method in machine learning to show a sequence of inter-related features/attributes and targets [10]. These relations represented by a tree where each node represents a feature/attribute. The tree has a root node in the top as the best predictor, branches in the middle as the possible choices and leafs node in the bottom as the decisions. From this representation, a set of rule can be used to classify/predict the decision. Thus, the simple way of Decision Tree which mimic the human thinking is making the decision process easy to understand (not like blackbox algorithm such as deep learning, SVM, etc.).

There are many types of decision trees like ID3 [11] and C4.5, CART, the probabilistic decision tree [12], the fuzzy decision tree [13] and the random forest [14]. ID3 (Interactive Dichotomizer 3) is an algorithm to build a decision tree using entropy. Entropy is used to select the best attribute in each step of the growth tree. C4.5 is extended algorithm of ID3. The difference between them is only on the number of split in the tree. If ID3 uses binary split, C4.5 uses multiple split. CART (Classification and Regression Tree) is another algorithm for decision tree. It uses Gini coefficient to determine which attribute will be splitted [15]. The decisions usually come with uncertainty. Thus, Probabilistic Decision Tree and Fuzzy Decision Tree are coming up to handle the crisp decisions. Random forest is a kind of decision tree which build many trees using CART algorithm, and it classifies new instance by using majority vote. Further study should be conducted to find out which one is better for flower's recognition problem.

BN is another approach that seems also have a possibility to solve these problems. Since it represented by a structure which have directed arch, it can be easy to explain the decision process. BN is different with the tree. If the tree only has a root, BN can have multiple roots. The tree reflects how the attributes can affect the target, without care about the dependency. Meanwhile, BN is very concern to the direction of arch. It refers to conditional independency that has a big impact in the decision. In BN, it is not permitted to get a cyclic graph. Thus, there is a specific method and pattern to get BN structure. Not like Decision Tree which only built from data, BN structure provides an alternative by combining data and expert knowledge when there is not sufficient data [16]. By using BN, we can know the conditional dependencies between the variables, compute the joint probability table, and determine the probability of the non-evidence variables given the evidence.

2.2 Integrating taxonomist knowledge to the automated image-based system

In this research, we will first focus on images of flowers because an identification based on the complete plant would be too complex. To find a bridge between what taxonomists do in traditional identification and what computer scientists do in image processing, we will acquire the characteristics of flowers that are usually used by taxonomists. When taxonomists identify a flower, they will start by checking, for example, the types of flower arrangement (inflorescence). Then, they also check the symmetry, shape, color, texture and other characteristics of the flower. Figure 1 is an illustration of how this could work. This tree is not static because the taxonomist can proceed in a different way.

From those information, we can adopt taxonomists work by making the structure that can explain and find the decision using the architecture explained in Section 2.1. By using the proposed architectures, we can integrate taxonomist knowledge into the system.

Moreover, one of the basic characteristics of the proposed architecture is feature/ attribute by node. We note that we need a classifier to deal with properties of the plant at that level. So, part of the research is how to design this classifier. Every node in this tree is a classifier that we have to design, and the information is to be extracted from the images. For example, nodes 1-6 in Figure 1 are the classifiers for checking: 1) the type of inflorescence, 2) the type of cluster, 3) the symmetry of the flower, 4) the labellum, 5) the shape of the flower, 6) color and texture.

2.3 Handling Uncertainty

Even though the proposed architectures have mechanism to compute the probability of a final result, but the uncertainty does not only come from the result. For example, to check the type of inflorescence in Figure 1. If after read the image, the system does not sure whether the flower is single or cluster, then the system can give us the probability 0.5 for single and 0.5 for cluster. Another example is determine the colour. Sometimes we say if the flower has blue colour. But, another person will say if it has purple colour.

To handle this uncertainty, we need a classifier that can handle the probability of each attribute. It will not be easy because we directly extract the flower characteristic from the image and determine its probability. This mechanism will apply to our proposed architecture and the trustworthiness of the system will improve.

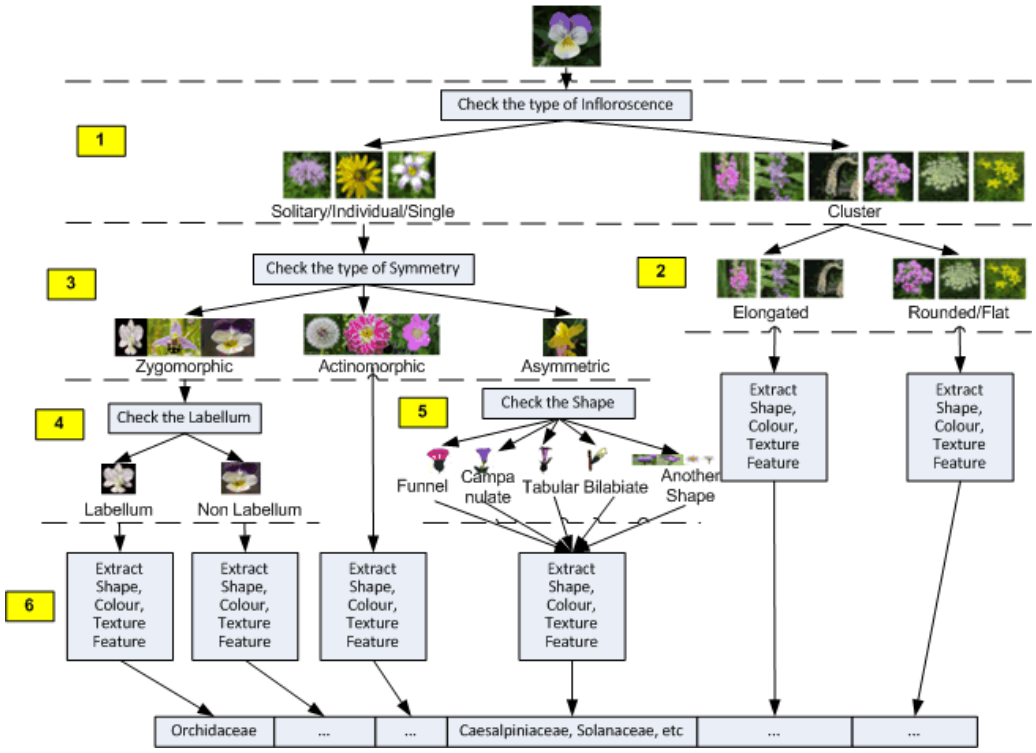


Figure 1: Flower Properties

3 Experiment and Preliminary Result

The starting point of our approach is the decision tree. So, in this paper we conducted a little experiment to implement the new perspective in plant recognition using that approach. This experiment does not cover all of our proposed method. We just make a simulation with a small data to know whether the decision tree can be used in our case. Firstly, we collect the information about the flower characteristics from the online flower identification sofware (GoBotany [6]). This step is used temporarily to substitute reading characteristics from the image. After our design classifier is ready to use, for the next, the characteristic should be read directly from the image and become the input of the decision tree.

We used 3 attributes and 5 species of orchid flower. The attributes are colour of flower, colour of labellum and also texture, while the species are *Arethusa bulbosa, Calopogon tuberosus, Corallorhiza maculata, Corallorhiza trifida, and Cypripedium acaule.* For the simplicity, we symbolize the species by A, B, C, D, and E sequentially. From the information that we got, then we generate the synthetic data. The number of samples that we generated are 125. Some of the samples can be looked at Figure 2. As the first attempt, we used CART algorithm to build the tree from those data. The decision tree yielded by the algorithm can be shown in Figure 3.

Based on the decision tree in Figure 3, we can predict the species of the new data. For example, if we have the flower with these caharacteristics: colour of flower is white, colour of labellum is yellow, and it has spot, then the decision tree will decide D for the species. To test the performance of the decision tree and avoid overfitting, we used cross validation with k-fold=5. The overall performance of the decision tree in this case is still low. The accuracy of the system is 64 %. It can show by the confusion matrix in Figure 4 where species C is the most difficult species to identify. The accuracy itself is affected by some components like the number of samples, the features that we used,

Editor - tigapuluh.m

yfit | fishertable | T

125x4 table

	1 CF	2 CL	3 Class	4 T
1	'Purple'	'Pinktored'	'C'	'Spot'
2	'Pink'	'White'	'C'	'Spot'
3	'Purple'	'Pinktored'	'A'	'Spot'
4	'Green'	'Pinktored'	'B'	'NoSpot'
5	'Blue'	'Pinktored'	'A'	'Spot'
6	'Yellow'	'White'	'C'	'Spot'
7	'White'	'Pinktored'	'B'	'NoSpot'
8	'Brown'	'Pinktored'	'A'	'Spot'
9	'Pink'	'Pinktored'	'A'	'Spot'
10	'Blue'	'White'	'B'	'NoSpot'
11	'White'	'White'	'B'	'NoSpot'
12	'Purple'	'Pinktored'	'E'	'NoSpot'
13	'Green'	'Pinktored'	'A'	'Spot'

Figure 2: The Samples

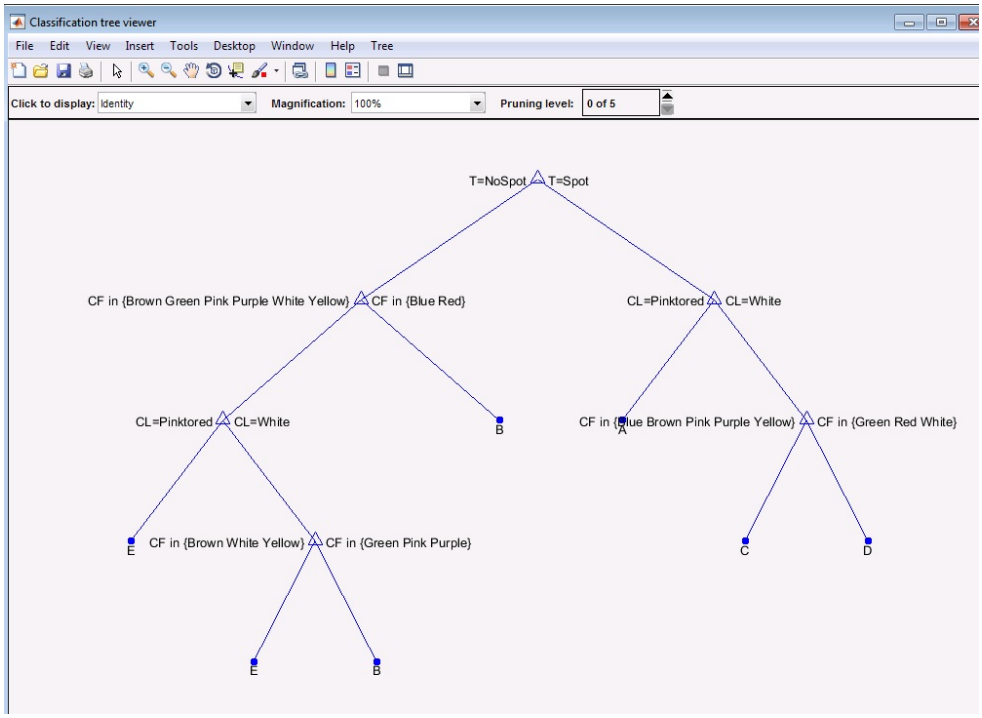


Figure 3: The Decision Tree

and also the algorithm that we choose. It needs more experiments about that.

Currently, our focus is not only in the accuracy of the system, but also on how to assign the probability in the decision tree. How if the color of the labellum is between yellow and orange, and the color of flower is not fully white, maybe like white greyish or white yellowish? The experiment about this issue have not implemented yet. Once more homework, in the next research this issue should be handled.

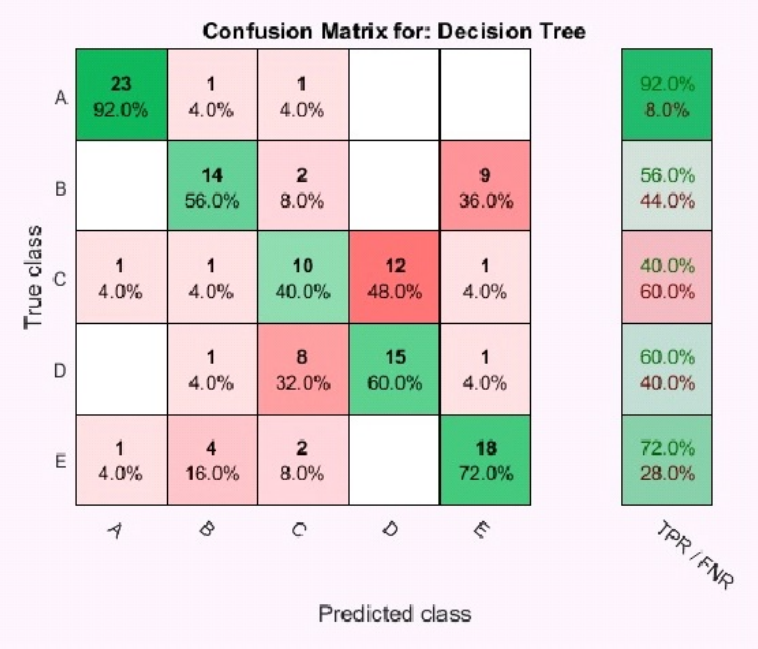


Figure 4: Confusion Matrix

4 Conclusion

We have proposed a new perspective in automated image-based flower recognition by designing a system that acts not as a blackbox system. From the experiment we have conducted, the decision yielded by the decision tree is quite low, with 64 % accuracy. Further research about implementation using the decision tree is still needed. Besides that this proposed method needs to be implemented as a whole and compares to another architecture in order to get the best performance.

Acknowledgment

The research described in this paper was supported by Research and Innovation in Science and Technology Project (RISET-Pro) of Ministry of Research, Technology, and Higher Education of Republic Indonesia (World Bank Loan No.8245-ID).

References

[1] https://prasetya.ub.ac.id/berita/Prof-Darnaedi-Indonesia-Langka-Ahli-Taksonomi-12422-id.html, last accessed on Oct 1, 2017.

[2] Gaston KJ, ONeill MA, "Automated species identification: why not?" Philos Trans R Soc Lond, B Biol Sci 359(1444):655667, doi:10.1098/rstb.2003.1442, 2004.

[3] Waldchen J, Mader P, "Plant species identification using computer vision techniques: A systematic literature review", Arch Computat Methods Eng., doi:10.1007/s11831-016-9206-z, 2017.

[4] Watson, L., and Dallwitz, M.J, "The families of flowering plants: descriptions, illustrations, identification, and information retrieval", Version: 30th September 2017, 1992 onwards.

[5] Glenny D, James T, Cruickshank J, Dawson M, Ford K, Breitwieser I, "Key to flowering plant genera of New Zealand", Accessed at http://www.landcareresearch.co.nz/resources/identification/plants/flowering-plants-key, 2012.

[6] https://gobotany.newenglandwild.org/full/, last accessed on Oct 11, 2017.

[7] http://www.colby.edu/info.tech/BI211/, last accessed on Oct 11, 2017.

[8] http://hort.ifas.ufl.edu/floragator/, last accessed on Oct 11, 2017.

[9] http://kukkakasvit.luontoportti.fi/index.phtml?lang=en, last accessed on Oct 11, 2017.

[10] Lucey, T. and Lucey, T., "Quantitative Techniques", 6th Edition, Book Power, London, 2002.

[11] Quinlan, J.R., "Induction of Decision Trees", Machine Learning , 1, Kluwer Academic Publishers, 81-106, 1986.

[12] Quinlan, J. R., "Probabilistic decision trees In Machine learning", Yves Kodratoff and Ryszard, S. Michalski (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA 140-152, 1990.

[13] T. C. Wang, and H. D. Lee, "Constructing a Fuzzy Decision Tree by Integrating Fuzzy Sets and Entropy", WSEAS Transactions on Information Science and Applications, vol. 3, no. 8, pp. 1547-1552, 2006.

[14] Breiman, L, "Random Forest", Machine Learning, 45, p. 5-32, 2001.

[15] Songul, C., "Comparison of Performance of Decision Tree Algorithms and Random Forest: An Application on OECD Health Expenditures", International Journal of Computer Applications (0975-8887), Volume 138, No. 1, March 2016.

[16] sucar, L.E. "Probabilistic Graphical Models: Principles and Applications",Springer Publishing Company, Incorporated, 2015.

Particle Filter-based Parameter Estimation in a Model of the Human Circadian Rhythm

Jochem H. Bonarius[1] Jean-Paul M.G. Linnartz[1,2]
[1]Eindhoven University of Technology [2]Philips Lighting Research
Electrical Engineering, Signal Processing Systems
FLX 7.062, Postbox 513, 5600 MB, Eindhoven, The Netherlands
j.h.bonarius@tue.nl j.p.linnartz@lighting.com

Abstract

Recent insights into the effects of light on human health call for a more human-centric approach in automatic lighting control systems. We contribute to the provisioning of lighting settings tailored to the needs of individuals by addressing the challenge of predicting the response of an individual's circadian rhythm to light exposure. Existing models of the human circadian rhythm are not tailored to individual physiological characteristics such as intrinsic circadian period, light sensitivity and age. We propose to improve model accuracy by using Bayesian statistical inference to estimate the values of model parameters that reflect these physiological characteristics. We illustrate our generic method by applying to a combination of two popular models of the circadian rhythm. By processing individual light exposure- and actigraphy data recoded during a field trial with 20 human subjects with a Particle Filter, we estimate each subject's intrinsic circadian period. When correlating these to the subjects' Munich Chronotype Questionnaire Midsleep on Free Days time, a significant relationship was found: $r > 0.6$ and $p < 0.01$. This shows the proposed method has good potential for improving model accuracy.

1 Introduction

Humans have an internal circadian rhythm that regulates many of their biological processes such as temperature, hormone secretion, and the sleep-wake cycle. The timing of light exposure plays a major role in regulation of this circadian rhythm [1]. Determining the state of the circadian cycle has been a major subject in the field of Chronobiology. Within this field several mathematical models of the circadian rhythm have been proposed, often based on empirical observations gathered in clinical studies. Commonly, the human circadian rhythm is modeled as a deterministic system with certain inputs (light exposure, food intake, etc.) and outputs (body temperature, social markers, etc.). However, as these models were often created by fitting mathematical functions to the average of the collected data, their output will represent the average response, not that of an individual. This could lead to misprediction, for example were the model would indicate a circadian phase advance in response to certain light exposure, while actually the individual's circadian phase would be delayed.

We propose to improve model accuracy by using Bayesian statistical inference to estimate the value of model parameters that reflect physiological characteristics such as intrinsic circadian period, light sensitivity and age. Not only do these differ per individual, but they are also not always fully known. By observing an individual's

responses to inputs, we can iteratively update the estimation of the parameter values. This is schematically shown in figure 1.

Our target is to estimate parameter values that best correspond to an individual's characteristics, in order to reduce the modeling error for that individual. As we want to implement our models in automatic lighting control systems, we do not aim for a clinically accurate estimate, but we need an estimate that is adequate for choosing between different options for light settings.

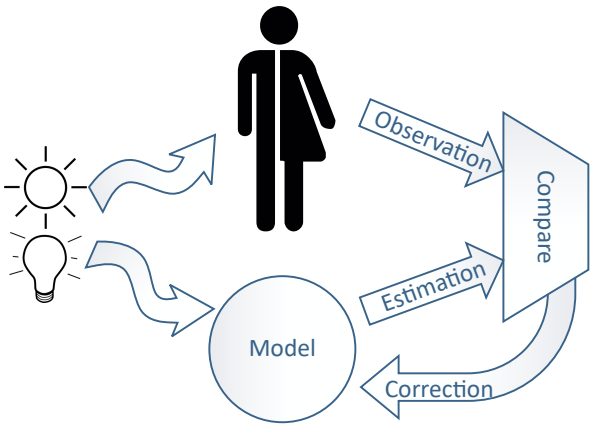


Figure 1: Schematic representation of the model parameter update loop. Unisex symbol ©Scott de Jonge

The use of Bayesian inference in this context has already been suggested by Mott, Dumont, Boivin, *et al.* [2]. They showed how a Particle Filter, a Sequential Monte Carlo method, can be used to approximate the system state (circadian phase) by observing light exposure and body temperature. We will extend on this method to estimate the models' parameter values using techniques developed by Liu and West [3].

2 Methods

2.1 Parameter search using a particle filter

We consider the circadian rhythm to be Markov process with transition density $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \boldsymbol{\theta})$ and observation density $p(\mathbf{z}_k|\mathbf{x}_k, \boldsymbol{\theta})$. We want to determine the probability distribution of a (sub)set of fixed parameters in vector $\boldsymbol{\theta}$, given all observations $\mathbf{z}$ up to now: $p(\boldsymbol{\theta}|\mathbf{z}_{1:k})$. Using Bayes' theorem, the Chapman–Kolmogorov equation, and by including the state variable $\mathbf{x}$, the probability can be rewritten as an iterative algorithm, where the current state $\mathbf{x}_k$ and parameter estimation $\hat{\boldsymbol{\theta}}_k$ depend on the previous, according to

$$p(\hat{\boldsymbol{\theta}}_k, \mathbf{x}_k|\mathbf{z}_{1:k}) \propto \int p(\hat{\boldsymbol{\theta}}_k|\hat{\boldsymbol{\theta}}_{k-1}, \mathbf{x}_k, \mathbf{z}_k) p(\mathbf{z}_k|\hat{\boldsymbol{\theta}}_{k-1}, \mathbf{x}_k) \int p(\mathbf{x}_k|\hat{\boldsymbol{\theta}}_{k-1}, \mathbf{x}_{k-1}) p(\hat{\boldsymbol{\theta}}_{k-1}, \mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}) \, \mathrm{d}\mathbf{x}_{k-1} \, \mathrm{d}\hat{\boldsymbol{\theta}}_{k-1}, \quad (1)$$

By defining the term $p(\hat{\boldsymbol{\theta}}_k|\hat{\boldsymbol{\theta}}_{k-1}, \mathbf{x}_k, \mathbf{z}_k)$, where the new estimation of $\boldsymbol{\theta}$ only depends on the previous -state, -observation, and parameter estimation and not on their entire history, it is implied that $\boldsymbol{\theta}$ has a simple, known distribution. Liu and West [3] suggest that this parameter density can be approximated using a weighted kernel

density constructed by adding N multivariate Gaussian densities*

$$p(\hat{\boldsymbol{\theta}}_k|\mathbf{x}_k, \mathbf{z}_k, \hat{\boldsymbol{\theta}}_{k-1}) \approx \sum_{i=1}^{N} w_k^{(i)} \mathcal{N}_{\dim(\boldsymbol{\theta})}(\hat{\boldsymbol{\theta}}_k|\mathbf{m}_k^{(i)}, (1-a^2)\mathbf{V}_k), \tag{2}$$

which was further reduced to

$$\hat{\boldsymbol{\theta}}_k^{(i)} \sim \mathcal{N}_{\dim(\boldsymbol{\theta})}(\mathbf{m}_k^{(i)}, (1-a^2)\mathbf{V}_k), \text{ for } i = 1, 2, \ldots, N. \tag{3}$$

'Smoothed' Gaussian mean vector $\mathbf{m}$ in the previous equations is determined using a mixture of the previous parameter estimate $\hat{\boldsymbol{\theta}}$ and the posterior parameter mean $\bar{\boldsymbol{\theta}}$

$$\mathbf{m}_k^{(i)} = a\hat{\boldsymbol{\theta}}_{k-1}^{(i)} + (1-a)\bar{\boldsymbol{\theta}}_k, \text{ for } i = 1, 2, \ldots, N, \tag{4}$$

where smoothing factor a (also in equation 2) is determined according to

$$a = \frac{3\delta - 1}{2\delta}, \tag{5}$$

for which we use a fixed discount factor $\delta = 0.98$. The posterior parameter mean $\bar{\boldsymbol{\theta}}$ is determined by

$$\bar{\boldsymbol{\theta}}_k = \sum_{i=1}^{N} w_k^{(i)} \hat{\boldsymbol{\theta}}_{k-1}^{(i)}. \tag{6}$$

The (normalized) particle weight $w^{(i)}$ is derived from the observation density, described by

$$w_k^{(i)} = \frac{p(\mathbf{z}_k|\boldsymbol{\mu}_k^{(i)}, \hat{\boldsymbol{\theta}}_{k-1}^{(i)})}{\sum_{i=1}^{N} p(\mathbf{z}_k|\boldsymbol{\mu}_k^{(j)}, \hat{\boldsymbol{\theta}}_{k-1}^{(j)})}, \text{ for } i = 1, 2, \ldots, N. \tag{7}$$

Here, the mean value $\boldsymbol{\mu}$ of the state $\mathbf{x}$ is determined by determining the expected value of the state equation (defined later-on in this paper) with

$$\boldsymbol{\mu}_k^{(i)} = \mathbb{E}\left[\mathbf{x}_k \middle| \mathbf{x}_{k-1}^{(i)}, \hat{\boldsymbol{\theta}}_{k-1}^{(i)}\right], \text{ for } i = 1, 2, \ldots, N. \tag{8}$$

The posterior covariance matrix of the parameter distribution $\mathbf{V}$ is described by

$$\mathbf{V}_k = \sum_{i=1}^{N} w_k^{(i)} (\hat{\boldsymbol{\theta}}_{k-1}^{(i)} - \bar{\boldsymbol{\theta}}_k)(\hat{\boldsymbol{\theta}}_{k-1}^{(i)} - \bar{\boldsymbol{\theta}}_k)^{\mathrm{T}}. \tag{9}$$

Equation 2 implies a point-mass representation can be used to approximate the parameter density, which we realize using a particle filter [4]. Hence, in the equations above, subscript (i) indicates the particle index and N represents the total number of particles. As the total number of particles is limited, it is important that the majority of particles provides an effective contribution to the point-mass representation. If the weight of most particles is close to zero, then the accuracy of the estimated probability distribution will be low. To prevent this from happening, a *resampling* step is used after each iteration: particles with low weight are dropped and particles with high weight are replicated [4]. The new indexes for resampling are sampled from a multinomial distribution with parameters $p_i = w^{(i)}$, for $i = 1, 2, \ldots, N$. This will be shown in pseudocode at the end of this paper.

*We use $\mathcal{N}_D(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$ to denote the probability density function of a D-variate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $D = 1$ when omitted.

2.2 Mathematical model of the circadian rhythm

We consider the Kronauer limit cycle oscillator model, a popular mathematical model of the circadian pacemaker that estimates the response of the circadian pacemaker to ambient light input, as a base for our work [5]. However, the output markers that can be related to the state of that model, such as the time-of-minimum core body temperature ($\mathrm{CBT}_{\mathrm{min}}$) [5] or dim light melatonin onset (DLMO) [6], are impractical to measure in daily situations: a subjects either need to wear an internal thermometer, or lab analysis of saliva samples is required. Therefore, we combine the model with the mathematical model of the homeostatic sleep drive by Phillips and Robinson [7] that relates the circadian clock to the sleep-wake cycle, as its relatively easy to use actigraphy to determine the sleep-wake state of an individual. In [8] and [9], it was already suggested to combine these models. Furthermore, the model combination also fits into the concept of the two-process model of sleep regulation [10].

We connect the models in a way that the system has an input vector $\mathbf{u}$ and an output vector $\mathbf{z}$, as shown schematically in figure 2.

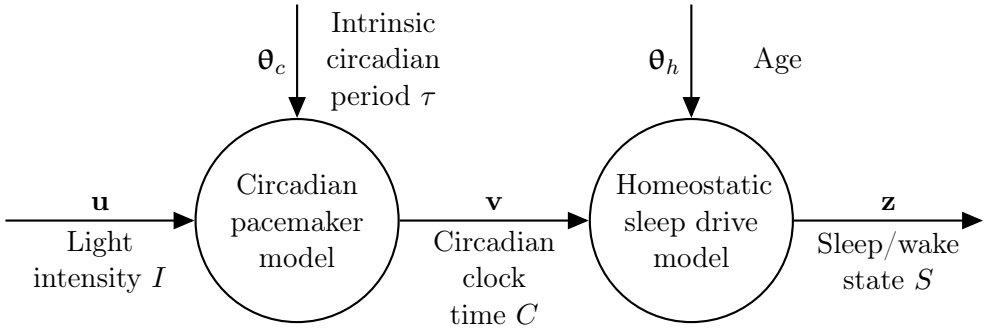


Figure 2: Block diagram of the interconnected models.

We combine all the models' system state variables in state vector $\mathbf{x} \triangleq \begin{bmatrix} n & x & y & V_v & V_m & H \end{bmatrix}^{\mathrm{T}}$. In this work only the parameter τ, representing the period of the circadian pacemaker, is considered for estimation, as there are strong indications that this parameter is the dominant source behind individual variation in circadian phase [6]. We consider all other parameters fixed at their suggested value. The dynamics of the system are then described by

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{n} \\ \dot{x} \\ \dot{y} \\ \dot{V}_v \\ \dot{V}_m \\ \dot{H} \end{bmatrix} = \begin{bmatrix} 60(\alpha(1-n) - 0.007n) \\ \frac{\pi}{12}\left(y + 0.13\left(\frac{1}{3}x + \frac{4}{3}x^3 - \frac{256}{105}x^7\right) + B\right) \\ \frac{\pi}{12}\left(\frac{1}{3}By - x\left(\left(\frac{24}{0.99729\tau}\right)^2 + 0.55B\right)\right) \\ 360\left(D_v - V_v - 2.1Q_m\right) \\ 360\left(1.3 - V_m - 1.8Q_v\right) \\ \frac{1}{45}\left(\mu_H Q_m - H\right) \end{bmatrix} \triangleq \mathbf{f}(\mathbf{x}, I; \tau), \tag{10}$$

describing dynamics of the ratio of activated photoreceptors n, the circadian pacemaker oscillator pair x and y, the mean cell body potential of the sleep-active ventrolateral preoptic (VLPO) area of the hypothalamus V_v, the mean cell body potential of the wake-active ascending arousal system's monoaminergic nuclei (MA) V_m, and the

homeostatic sleep drive H. Supporting equations are

$$\alpha = 0.1\sqrt{\frac{I}{9500}}\frac{I}{I+100}, \tag{11}$$

$$B = 37\alpha(1-n)(1-0.4x)(1-0.4y), \tag{12}$$

$$C = 0.5(1+0.8x-0.55y), \tag{13}$$

$$Q_i = \frac{100}{1+\exp\left(\frac{10-V_i}{3}\right)}, \text{ with } i \in \{v,m\}, \text{ and} \tag{14}$$

$$D_v = H - \nu_{vc}C - 10.2, \tag{15}$$

describing the photoreceptor activation rate α following light exposure I, the resulting photic drive B, the circadian clock time C (modified from [8] to fit the model from [5]), the mean firing rate of the VLPO Q_v and MA Q_m and the drive on the VLPO mean cell body potential D_v. In the equations, homeostatic dampening factor μ_H and circadian clock sensitivity ν_{vc} are related to the age of the modeled person [8]. We will suggest values for these variables based on our data set in the Results section.

As output, only the sleep-wake state S_{sw} is considered, because it is relatively easy to measure in ambulatory conditions as was explained at the beginning of this subsection. It is derived from [7] as[†]

$$S_{sw} \triangleq \mathcal{H}(Q_m - 1) = \begin{cases} 1(\text{awake}), & \text{if } Q_m \geq 1 \\ 0(\text{sleeping}), & \text{otherwise} \end{cases}. \tag{16}$$

However, the above equation is constant most of the time: every circadian cycle ($\sim$ 24 h) only one 0-to-1 transition (wake up) and one 1-to-0 transition (sleep onset) occurs[‡]. The times in-between transitions do not give us much information. Therefore, only the transitions are considered interesting for our parameter estimation. Thus, we introduce observation set Z which contains all the times t at which a 0-to-1 or 1-to-0 transition occurs in S_{sw} - that is, the transition times from sleep to wake or vice versa. We then evaluate the system output $z_k \in Z$, for $k = 1, 2, \ldots, 2 \times \#\text{days}$, i.e. two events per day. Effectively, we evaluate the model until a $Q_m = 1$ event occurs, indicating either a sleep onset or a wake up time. We will then compare this estimated time (denoted as $\hat{z}$) with the actual sleep onset or wake up time z observed with the human subject. To support this, we define output function $\mathbf{h}$ which maps state $\mathbf{x}$ to observation z by evaluating equations 14 and 16 and determining the time a transition occurs.

As the above equations show, the models of the circadian pacemaker and the homeostatic sleep drive are described to be deterministic. However, real-life biological processes are stochastic in nature. We introduce stochasticity into the existing model by adding white Gaussian process noise and -measurement noise to the state respectively

[†]We use $\mathcal{H}(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0 \end{cases}$ to denote the unit-/Heaviside step function.

[‡]In reality the sleep cycle is much more complex and a person can actually wake up multiple times during that cycle. But our simplified model only considers the initial sleep onset and the final wake up time.

the output. Following this, the state transition density and observation density are described by

$$p\left(\mathbf{x}_k|\mathbf{x}_{k-1},\boldsymbol{\theta}\right) = \mathcal{N}_6\left(\mathbf{x}_k|\mathbf{F}_k(\mathbf{x}_{k-1},I;\tau),\boldsymbol{\Sigma}_\mathbf{x}\right) \text{ and} \tag{17}$$

$$p\left(\mathbf{z}_k|\mathbf{x}_k,\boldsymbol{\theta}\right) = \mathcal{N}\left(z_k|\mathbf{h}(\mathbf{x}_k),\Sigma_z\right), \tag{18}$$

with $\mathbf{F}_k$ being the discrete-time approximation of the state equation 10. Since the mean of additive white Gaussian noise is 0, only the covariance matrix of the process noise $\boldsymbol{\Sigma}_\mathbf{x}$ and the variance of the measurement noise Σ_z appear in the equations. Determining the process noise in covariance matrix $\boldsymbol{\Sigma}_\mathbf{x}$ is out of the scope of this work. For now we assume a value of $\boldsymbol{\Sigma}_\mathbf{x} = (0.01)^2\mathbf{I}_6$. Next, we assume that the variance of the observation noise Σ_z is related to the variance in sleep-onset and wake-up times. Therefore, we set Σ_z equal to the variance of the sleep onset- and wake up times observed with the human subject under evaluation.

With the state transition density and observation density defined, the particle filter described in subsection 2.1 can be constructed [4]. Determining the optimal number of particles is out of the scope of this research. Instead, $N = 240$ particles, suggested by Mott, Dumont, Boivin, *et al.* [2], is used as it shows consistent results.

The particle filter was implemented in MATLAB. We approach $\mathbf{F}_k$ numerically using MATLAB's ordinary differential equation solver "ode23s". A pseudo-code description of the particle filter implementation can be found in Algorithm 1 at end of this paper.

3 Results

To illustrate our method, the particle filter algorithm was applied to data obtained in a field study with 20 human subjects. However, 4 data sets had to be dropped because of hardware issues and user errors. For the 16 data sets left, the average age of the subjects was $70.9 \pm 4.0\,\mathrm{yr}$.

Each subject wore a Philips Actiwatch Spectrum Pro, measuring actigraphy, and a Martin light-logger, measuring ambient light intensity, for a minimum of 168 h (7 days). In parallel, each subject was asked to maintain a sleep dairy, indicating their "to bed"- and "out of bed" times. As described by the Munich Chronotype Questionaire (MCTQ) [11], this information can be used to determine the subjects' sleep preference (Chronotype). As the subjects are retired and don't use alarm clocks, their sleep preference (Chronotype) can be directly derived from their Midsleep on Free Days time (MSF), as described by

$$\mathrm{MSF} \triangleq 0.5\left(t_{\text{sleep onset}} + t_{\text{wake up}} - 24\,\mathrm{h}\right). \tag{19}$$

By combining the sleep diary data with actigraphy data recorded by the Actiwatches, each subject's sleep onset and wake up times were estimated by hand to form observation set Z. For example, for subject 17, the (partially shown) set is

$$Z = \left\{01{:}15,\quad 08{:}05,\quad 26{:}05,\quad 32{:}50,\quad 49{:}40,\quad \ldots,\quad 152{:}00\right\}. \tag{20}$$

In [8], the values for age-related parameters ν_{vc} and μ_H are suggested to be $\nu_{vc} \approx 2.35\,\mathrm{mV}$ and $\mu_H \approx 3.95\,\mathrm{nM\,s}$ for old age. However, analysis of the data showed that

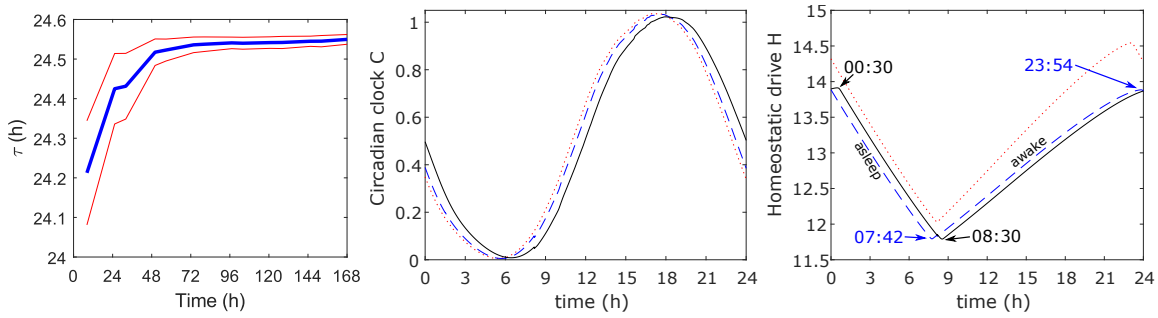


Figure 3: Particle filter output for subject 17 (male, 73 yr). The first graph shows the estimated intrinsic circadian period τ as a function of simulation time. The blue middle line shows the posterior mean and the red lines show the posterior standard deviation. The graph shows that τ converges to 24.55 h with an exponential curve. The second and third graphs show the circadian clock time C and homeostatic sleep drive H at day 7 of the collected data. Red dotted line: the original models with their original parameters. Blue dashed line: a particle filter(PF) with proposed parameters $\nu_{vc} = 2.9\,\mathrm{mV}$ and $\mu_H = 4.0\,\mathrm{nM\,s}$, only estimating the state. Black solid line: the proposed PF, estimating state and τ. The state-only PF estimates a small circadian phase delay of ~ 20 min, while the proposed PF estimates a more significant delay of ~ 60 min. This is reflected in the homeostatic sleep drive output: the estimated sleep onset/wake up times of the proposed PF are closer to the actual times observed with this subject: sleep onset at 01:29 and wake up at 08:12.

$\nu_{vc} = 2.9\,\mathrm{mV}$ and $\mu_H = 4.0\,\mathrm{nM\,s}$ best fit our data set, which we will use in our experiments. Further analysis of the data showed that the average initial state $\bar{\mathbf{x}}_0 = [\ 0.25\ -0.9\ -0.5\ 2.5\ -12\ 13.8\ ]^{\mathrm{T}}$.

The light data of each subject was individually processed by the particle filter, using the sleep-wake times as observation input. To illustrate the results, the proposed Particle Filter's output for subject 17 is shown and compared to prior methods in figure 3.

In [12], the MCTQ MSF has been associated with the intrinsic period of the circadian pacemaker τ. Therefore, we correlate the resulting posterior mean of the intrinsic period for each subject to that subject's MSF time using linear regression analysis. The results of two successive runs can be seen in figure 4. The Pearson correlation coefficient shows significant correlation with strength $r > 0.6$ and significance $p < 0.01$, which indicates that our proposed method can estimate the intrinsic period of the circadian pacemaker.

4 Discussion and Future Work

Our proposed method utilizes both input- (light exposure) and output (observation) data, which both contain information about the circadian rhythm of an individual. In our study we specifically use the observation set Z that contains natural sleep onset and wake up times of an individual from which we extract information indicating the actual circadian phase for this individual. At the same time, we also use this data to determine the subject's MCTQ MSF. This works well for our situation. However, the natural sleep-wake rhythm is disrupted in the case an individual uses an alarm clock: in that case we lose an important observation channel. In a test with a second dataset where the subjects were using alarm clocks, we did not find a significant correlation,

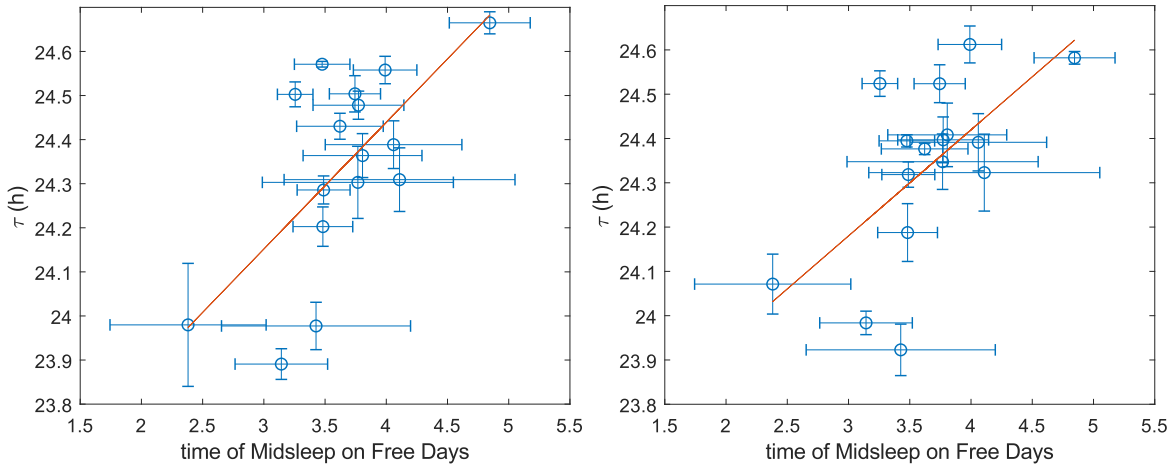


Figure 4: Two scatter plots showing the output of two sequential runs of the particle filter with the data from the study. The estimated intrinsic circadian period τ on the vertical axis is plotted against the MCTQ MSF time on the horizontal axis. For each of the estimated values, the standard deviation is shown as an error bar. The linear regression line is shown in red. The output of the sequential runs show that the particle filter will give a different result every run. This is caused by the stochasticity the underlying model. However, both runs show the correlation is significant: Pearson's $r = 0.66$ with $p = 0.005$ for the left plot and $r = 0.62$ with $p = 0.0099$ for the right plot.

which is intuitively appealing. It is a subject of our future research to find a suitable alternative input or output parameter that provides information on the circadian phase.

Our study shows that one can get a good estimation of model parameters even with a very limited number of particles. Adding more particles does not improve the results. This is surprising because we consider relatively many variables: In Mott, Dumont, Boivin, *et al.* [2] only the Kronauer model with 3 state variables is used. By including the Phillips and Robinson model and searching for τ, we add 4 more variables. This would suggest $(240)^{7/3} \approx 360000$ particles are needed. We believe that we can work with fewer particles because our initial state is very close to the actual state, because the homeostat model closely follows the circadian clock, and because the process noise (in $\mathbf{\Sigma_x}$) is chosen quite small. Hence, our particle filter barely has to put effort in finding the state. Most effort goes into finding τ, which is feasible with only a small number of particles. In further studies, we want to explore to what extent increased process noise degrades the particle filter results, or could even cause divergence. In such case more particles would be required.

Our tests revealed a statistical deviation of the parameters ν_{vc} and μ_H from the age-dependent model suggested by [8]. In fact, we saw difference between individuals of the same age. This can explain the mismatch between estimated- and actual sleep-wake times for instance shown in figure 3. Our results suggest that these parameters should preferably also be estimated for each individual and therefore should be included in $\boldsymbol{\theta}$.

The average estimated intrinsic circadian period τ for all participants in our population is around 24.4 h. This is notably higher then the mean of 24.18 h determined by Czeisler, Duffy, Shanahan, *et al.* [13] and comparable research. This can be coincidently related to the selection of our participants.

The time interval $\Delta t = z_k - z_{k-1}$ is not constant. As a result the process- and observation noise covariance matrices $\mathbf{\Sigma_x}$ and Σ_z depend on k. However, because under

normal conditions the expected time between two sleep onset or wake up times is 24 hours ($\mathbb{E}[z_k - z_{k-2}] = 24\,\mathrm{h}$), it is reasonable to assume the modeled noise is sufficiently accurate.

5 Conclusion

Existing mathematical models of the circadian rhythm often model the average response of physiological processes that control the human circadian rhythm, which does not represent the response for an individual. This can result in a misprediction of an individual's circadian phase. We have shown how a Particle Filter can estimate values for the model's parameters to fit an individual's physiological characteristics. We have illustrated this by applying a Particle Filter to a combination of two existing models: the Jewett, Forger, and Kronauer circadian pacemaker model and the Phillips and Robinson homeostatic sleep drive model. By processing individual light exposure- and actigraphy data from 16 human subjects with the proposed Particle Filter, we estimate the parameter τ representing the subject's intrinsic circadian period. When correlating the estimated parameter values to the subjects' MCTQ MSF time, a significant relationship was found: $r > 0.6$ and $p < 0.01$. This demonstrates that a Particle Filter can estimate the intrinsic circadian period of an individual with reasonable accuracy, which will allow us to make a more accurate prediction of the effect that a specific lighting setting will have on the circadian rhythm of that individual.

6 Acknowledgments

We would like to thank Prof. Yvonne de Kort, Dr. Karin Smolders, Samantha Peeters, MSc. and Ruby van der Sande, MSc. of the Human-Technology Interaction group at the Eindhoven University of Technology for providing us with the data from the field study.

References

[1] T. A. Bedrosian and R. J. Nelson, "Timing of light exposure affects mood and brain circuits", *Transl. Psychiatry*, vol. 7, no. 1, 2017.

[2] C. Mott, G. Dumont, D. B. Boivin, and D. Mollicone, "Model-based human circadian phase estimation using a particle filter", *IEEE Trans. Biomed. Eng.*, vol. 58, no. 5, pp. 1325–1336, 2011.

[3] J. Liu and M. West, "Combined Parameter and State Estimation in Simulation-Based Filtering", in *Seq. Monte Carlo Methods Pract.* A. Doucet, Ed., Springer Science+Business Media New York, 2001, pp. 197–223.

[4] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/nongaussian bayesian tracking", *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, 2002.

[5] M. E. Jewett, D. B. Forger, and R. E. Kronauer, "Revised limit cycle oscillator model of human circadian pacemaker", *J. Biol. Rhythms*, vol. 14, no. 6, pp. 493–499, 1999.

[6] T. Woelders, D. G. M. Beersma, M. C. M. Gordijn, R. A. Hut, E. J. Wams, and A. Kota Gopalakrishna, "Daily Light Exposure Patterns Reveal Phase and Period of the Human Circadian Clock", *J. Biol. Rhythms*, vol. 32, no. 3, pp. 274–286, Jun. 2017.

[7] A. J. Phillips and P. A. Robinson, "A quantitative model of sleep-wake dynamics based on the physiology of the brainstem ascending arousal system", *J. Biol. Rhythms*, vol. 22, no. 2, pp. 167–179, 2007.

[8] A. C. Skeldon, A. J. K. Phillips, and D.-J. Dijk, "The effects of self-selected light-dark cycles and social constraints on human sleep and circadian timing: a modeling approach", *Sci. Rep.*, vol. 7, no. 45158, pp. 1–14, 2017.

[9] A. J. K. Phillips, C. A. Czeisler, and E. B. Klerman, "Revisiting Spontaneous Internal Desynchrony Using a Quantitative Model of Sleep Physiology", *J. Biol. Rhythms*, vol. 26, no. 5, pp. 441–453, 2011.

[10] A. A. Borbély, S. Daan, A. Wirz-Justice, and T. Deboer, "The two-process model of sleep regulation: A reappraisal", *J. Sleep Res.*, vol. 25, no. 2, pp. 131–143, 2016.

[11] T. Roenneberg, A. Wirz-Justice, and M. Merrow, "Life between Clocks: Daily Temporal Patterns of Human Chronotypes", *J. Biol. Rhythms*, vol. 18, no. 1, pp. 80–90, Feb. 2003.

[12] A. Hida, S. Kitamura, Y. Ohsawa, M. Enomoto, Y. Katayose, Y. Motomura, Y. Moriguchi, K. Nozaki, M. Watanabe, S. Aritake, S. Higuchi, M. Kato, Y. Kamei, S. Yamazaki, Y. I. Goto, M. Ikeda, and K. Mishima, "In vitro circadian period is associated with circadian/sleep preference", *Sci. Rep.*, vol. 3, pp. 1–7, 2013.

[13] C. A. Czeisler, J. F. Duffy, T. L. Shanahan, E. N. Brown, J. F. Mitchell, D. W. Rimmer, J. M. Ronda, E. J. Silva, J. S. Allan, J. S. Emens, D.-J. Dijk, and R. E. Kronauer, "Stability, Precision, and Near-24-Hour Period of the Human Circadian Pacemaker", *Science*, vol. 284, no. 5423, pp. 2177–2181, 1999.

Algorithm 1 Particle Filter for Intrinsic Circadian Period Estimation

Input: Light history I and sleep-wake times Z
Output: Posterior parameter mean: $\bar{\tau} = \sum_{i=1}^{N} w_k^{(i)} \hat{\tau}^{(i)}$

procedure ParticleFilter:
  for particle $i = 1, \ldots, N$ **do**
    Draw initial state: $\mathbf{x}_0^{(i)} \sim \mathcal{N}_6(\bar{\mathbf{x}}_0, \mathbf{\Sigma_x})$
    Draw initial theta: $\hat{\tau}_0^{(i)} \sim \mathcal{N}(24.18, (0.13)^2)$ (values from [13])
  end for
  foreach successive observation $z_k \in Z$ **do**
    for particle $i = 1, \ldots, N$ **do**
      Estimate mean state progression: $\boldsymbol{\mu}_k^{(i)} \leftarrow \mathbf{F}_k(\mathbf{x}_{k-1}^{(i)}, I; \hat{\tau}_{k-1}^{(i)})$
      Estimate observation: $\hat{z}_k^{(i)} \leftarrow \mathbf{h}(\boldsymbol{\mu}_k^{(i)})$
      Determine weight: $w_k^{(i)} \leftarrow \mathcal{N}(z_k | \hat{z}_k^{(i)}, \Sigma_z)$
    end for
    Normalize the weights: $w_k^i \leftarrow w_k^i / \sum_{j=1}^{N} w_k^j$
    Estimate posterior parameter mean: $\bar{\tau}_k = \sum_{i=1}^{N} w_k^{(i)} \hat{\tau}_{k-1}^{(i)}$
    Estimate posterior parameter covariance matrix: $V_k = \sum_{i=1}^{N} w_k^{(i)} (\hat{\tau}_{k-1}^{(i)} - \bar{\tau}_k)^2$
    for particle $i = 1, \ldots, N$ **do**
      Sample new index j from $\text{Multi}_N\left(w_k^{(1)}, w_k^{(2)}, \ldots, w_k^{(N)}\right)$
      Resample state: $\mathbf{x}_k^{(i)} \sim \mathcal{N}_6(\boldsymbol{\mu}_k^{(j)}, \mathbf{\Sigma_x})$
      Resample theta: $\hat{\tau}_k^{(i)} \sim \mathcal{N}(a\hat{\tau}_{k-1}^{(j)} + (1-a)\bar{\tau}_k, (1-a^2)V_k)$
    end for
  end for
end procedure

A Quantum Key Recycling Scheme based on qubits

Helena Bruyninckx
Royal Military Academy, Brussels, Belgium
helena.bruyninckx@rma.ac.be

Abstract

Quantum Key Recycling Schemes encode information into quantum states, such that the detection of eavesdropping is made possible. In case no eavesdropping was detected, the secret shared key can be safely re-used which is impossible for classical encryption schemes like the one-time pad. We propose a Quantum Key Recycling Scheme that encrypts a classical message into single BB84 qubits. Our scheme has a reduced key length compared to previous schemes.

1 Introduction

Quantum cryptography combines quantum theory with notions from classical cryptography in order to design secure quantum schemes. Typically, information is encoded in quantum states and the security is mainly derived by exploiting the counter-intuitive properties of those quantum states. Since handling quantum states is technically more challenging than handling classical bits, quantum schemes need to achieve a task that is not achievable by classical cryptography and/or need to add an extra functionality (eg, detect eavesdropping) [1].

In order to obtain classical information-theoretically secure encryption (or authentication), the secret key can be used only once even if no adversary was present. In the case of a quantum encryption (or authentication) scheme however, the key can safely be re-used since the participants can detect whether or not an adversary has tried to gain information about the exchanged quantum states. This idea of a Quantum Key Recycling Scheme was first proposed by Bennet, Brassard and Breidhart in 1982 [2]. Their scheme uses the same technology as what is needed for BB84 Quantum Key Distribution: it does not require quantum computers but only single-qubit operations. In [3], the authors propose a new scheme based on the same ideas as [2] and a scheme similar to [3] was proposed that significantly improved the noise tolerance [4].

In this short paper, we propose another QKRS having a reduced key length compared to [3]. A classical plaintext, together with a classical token, is encoded into BB84 quantum states by using a shared secret key. Verification is done by correctly measuring the quantum states, and comparing the result with the calculated information from the received message. In order to design the quantum scheme, several notions from classical cryptography are used, like hash-chains and universal hashing [5].

2 Overview of the scheme

We will discuss the problem of *message authentication* with key-recycling which can be extended to an encryption scheme as proposed in [3]. We denote $\mathcal{M}$ the set of classical messages, $\mathcal{T}$ the set of tags and $\mathcal{K}$ the set of shared keys. We consider classical plaintexts $msg \in \mathcal{M}$.

We use the principle of hash chains and the result of applying i times a hash function h ($h : \{0,1\}^* \to \{0,1\}^l$) to an input x_0 is denoted by $h^i(x_0)$. The counter i is called the *index* of the hash chain. This hash function is chosen uniformly at random from a *universal family of hash functions* $\mathcal{H}$ by means of a shared key $k \in \mathcal{K}$. Each family is indexed by a key k and consists of $2^{|k|}$ hash functions mapping a message $m \in \mathcal{M}$ into a l-bit hash output. Without knowing the key, an adversary can do no better than guessing which hash function was chosen. But even if she knows the hash values of a certain number of messages, this gives no information regarding the hash value of any other distinct message. This is because universal classes of hash functions behave like random functions with respect to collisions. To come up with another message-tag pair, an adversary can do no better than guessing.

The scheme also uses a Message Authentication Code $\mathsf{MAC} : \mathcal{K} \times \left(\mathcal{M} \times \{0,1\}^l\right) \to \mathcal{T}$. Consider a MAC with error probability $\epsilon_{\mathrm{MAC}} = 2^{-\lambda}$ and $|\mathcal{T}| = 2^{\lambda}$.

The classical bits are encoded into qubit states by using the 4-state encoding, also known as conjugate coding [6]. We write $|\psi\rangle = H^{\theta}|x\rangle$ to denote the plaintext $x \in \{0,1\}^n$ encrypted with key $\theta \in \{0,1\}^n$ where H denotes the Hadamard transformation.

The key material consists of a classical secret key $K_{AB} = k||K||\theta$. The key k is used to choose an element from the hash family, K is the MAC-key and θ denotes the basis sequence used for conjugate coding. The shared key material K_{AB} can be re-used for t rounds as long as no eavesdropping was detected. The counter t refers to the personal hash chain $h^t(K_A)$ constructed by Alice from a random $K_A \in \{0,1\}^*$. This will enable the authentication of t messages. We will use the short notation of K_A^t instead of $h^t(K_A)$.

We assume that the sender wants to send a classical message $msg \in \mathcal{M}$ to the receiver who knows the value of the counter i and K_A^t. The scheme works as follows

Authentication. The sender computes an element from her hash chain according to the counter i: $x = h^i(K_A)$. She computes a classical authentication tag $tag = \mathsf{MAC}_K(msg||x)$ and creates the quantum state $|\psi\rangle = H^{\theta}|tag||x\rangle$. Finally, she sends $|\psi\rangle$ and msg to the receiver.

Verification. Upon reception of $|\psi'\rangle$ and msg', the receiver measures the qubits $|\psi'\rangle$ according to θ and obtains the classical information tag' and x'. From x', he checks that $h^{t-i}(x') = K_A^t$. Then, he checks that $tag' = \mathsf{MAC}_K(msg||x')$ and outputs 0 or 1 accordingly.

Key-update. If the receiver accepts, re-use the shared keys K and θ. Otherwise, choose a uniformly random $\theta' \in \{0,1\}^n$ and update θ.

Noise in the quantum communication channel can be circumvented by adding a *secure sketch* [3] meaning that additional uniformly random secret key material is needed. Adding encryption of the message msg can be realized by an extractor that takes as input x and a seed. This seed is also part of the shared secret key between the sender and receiver. The extracted randomness is used as an one time pad to encrypt and decrypt the classical message msg. Since these last two steps are identical to the

steps in [3], we only focus on the differences between our authentication scheme and the Quantum Message Authentication Code QMAC from [3]. The same remark holds for the scheme from [4] since it also uses both ideas.

3 Discussion

Since handling quantum states is technically challenging, we focus on the key material needed for the quantum encryption. The other parts of the shared secret key (together with the secure sketch key and the extractor key) have the same length as in the scheme from [3]. In their Quantum Message Authentication Code QMAC, the sender first needs to choose a uniformly random $x \in \{0,1\}^r$ that is encoded into r qubits by means of a basis key θ. For each new message msg, another x needs to be chosen. The classical authentication tag $tag = \mathsf{MAC}_K(msg||x)$ is sent to the receiver without any encryption. The receiver verifies the correctness of the received message by measuring the qubits to obtain x and checking the authentication tag. For their scheme it is necessary that $r \geq 9\lambda - 1$ where $\lambda \in \mathbb{N}$ is the security parameter.

In our scheme, x is not chosen by the sender but was constructed from an initial secret by iteratively applying a hash function. Since this hash function is chosen uniformly at random from a family of hash functions, the result looks totally random. We encode our authentication tag, together with this x into n qubits. The value of n is related to the size of the authentication tag and the size of x, meaning: $n = \lambda + l$. Therefore, as long as $l < 8\lambda - 1$ our scheme uses less key material for the basis key θ than the scheme from [3].

We still need to consider the shared secret key k needed for the selection of the hash function, so this will induce a penalty of the size of k for the total amount of key material needed in our scheme.

In the QKRS from [4], additional key material is needed for an additional one time pad in order to protect the classical message authentication tag. This means that λ bits of key material are spent per round. For each authenticated message that the receiver accepts, this key needs to be replaced.

4 Conclusion

The Quantum Key Recycling Scheme that was proposed belongs to the family of QKRS that encodes classical information into qubit states. It follows a well known principle, more precisely adding redundancy in the message and than encoding everything into BB84 qubits. The description of the scheme focuses on the authentication problem and it allows to re-use the key when no eavesdropping was detected. Transforming this authentication scheme into an encryption scheme follows the same idea as [3]. The scheme uses less key material than other QKRS. Instead of the BB84 prepare and measure principle, we could use other encoding techniques like the 6-state or 8-state encoding [4].

References

[1] A. Broadbent and C. Schaffner, "Quantum cryptography beyond quantum key distribution," Design, Codes and Cryptography, January 2016.

[2] C.H. Bennett, G. Brassard and S. Breidbart. "Quantum cryptography II: How to re-use a one-time pad safely even if P =NP." In Natural Computing, 13(4):453 - 458 (2014). Original manuscript 1982.

[3] S. Fehr and L. Salvail. "Quantum Authentication and Encryption with Key Recycling. Or: How to Re-use a One-Time Pad Even if P =NP – Safely & Feasibly." In Advances in Cryptology, EUROCRYPT 2017, Springer-Verlag, 2017, pp. 311 - 338.

[4] B. Skoric and M. de Vries. "Quantum Key Recycling with eight-state encoding (The Quantum One Time Pad is more interesting than we thought)," 2016. https://eprint.iacr.org/2016/1122.

[5] N. Asokan, G. Tsudik, and M. Waidner, "Server-supported signatures," Journal of Computer Security, vol. 5, no. 1, pp. 91–108, 1997.

[6] S. Wiesner. "Conjugate coding," SIGACT News, 15 (1): 78-88, 1983.

Maximum Likelihood Decoding for Channels with Uniform Noise and Offset

Renfei Bu Jos Weber
Delft University of Technology
Applied Mathematics Dept., Optimization Group
Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands
R.Bu@tudelft.nl J.H.Weber@tudelft.nl

Abstract

In storage and communication systems, noise is not the only disturbance during data transmission. Sometimes the error performance can also be seriously degraded by offset and/or gain mismatch. This paper derives a maximum likelihood decoding criterion using intersection distance for channels with uniform noise and offset distribution in the absence of gain mismatch. Under this framework, zero word error rate performance is achievable for various decoding criteria by different standard deviation constraints on noise and offset. Our results show that the intersection distance decoding criterion incorporates the advantageous perks from two pre-existing methods: the immunity to offset mismatch that highlights Pearson distance decoding as well as the higher noise resistance brought by Euclidean distance decoding.

1 Introduction

We consider transmitting a codeword $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from a codebook $\mathcal{S}$, where n, the length of $\mathbf{x}$, is a positive integer. It is assumed that the received vector

$$\mathbf{r} = a(\mathbf{x} + \mathbf{v}) + b\mathbf{1}$$

is hampered not only by noise $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ but also by gain a and/or offset b, where $\mathbf{1}$ is the real all-one vector $(1, 1, \ldots, 1)$ of length n. The channels with gain and/or offset mismatch can commonly be found in storage and communication systems. Some examples are optical discs with fingerprints and scratches which may result in gain and offset variations of the retrieved signal [1], and charge leakage induced cell voltage shift in flash memory [2].

Traditional Euclidean distance decoding has been shown poor performances under such mismatches. Pearson distance decoding was proposed as an alternative to counter the effects of gain and/or offset mismatch [3]. Blackburn [4] investigated a maximum likelihood (ML) criterion for the channel with Gaussian noise and unknown gain and offset mismatch. In some applications, it is reasonable to assume a priori knowledge of the offset distribution. In [5], ML decision criteria are derived for Gaussian noise channels assuming the Gaussian or uniform distribution for the offset in the absence of gain mismatch. For Gaussian offset, the ML criterion turns out to be a weighted average of the Euclidean distance and the modified Pearson distance.

In this paper, we assume the absence of gain mismatch ($a = 1$), i.e.,

$$\mathbf{r} = \mathbf{x} + \mathbf{v} + b\mathbf{1},$$

as well as the uniform distribution of both the noise and the offset. For this channel model, we propose a maximum likelihood decoding criterion based on intersection distance (ISD), that possesses a property for countering both noise and offset mismatch.

Essentially, we combine the immunity to offset mismatch of Pearson distance decoding and the higher noise resistance of Euclidean distance decoding. We further show that, in this channel, zero word error rate (WER) performance is achievable by a small standard deviation of noise and/or offset.

The remainder of this paper is organized as follows. We start in Section 2 with a discussion of the prior art of decoding criteria. In Section 3, the ISD decoding criterion is presented. Zero WER performance is analyzed for Euclidean decoding, Pearson decoding, and intersection decoding in Section 4. Finally, we draw our conclusions in Section 5.

2 Preliminaries

For the noise vector $\mathbf{v} = (v_1, ..., v_n)$, assume the v_i are independently uniformly distributed with mean 0 and variance $\sigma^2 > 0$, i.e.,

$$\phi(\mathbf{v}) = \prod_{i=1}^{n} \phi(v_i) = \frac{1}{(2\sigma\sqrt{3})^n}, \ -\sigma\sqrt{3} < v_i < \sigma\sqrt{3}. \tag{1}$$

We assume that the offset b has a uniform probability density function ζ with mean μ and variance $\beta^2 > 0$. Since a receiver can subtract $\mu\mathbf{1}$ from $\mathbf{r}$ if the expected offset value is not equal to zero, we may assume $\mu = 0$ without loss of generality, which we will do throughout the rest of this paper. The probability density function of the offset is

$$\zeta(b) = \frac{1}{2\beta\sqrt{3}}, \ -\beta\sqrt{3} < b < \beta\sqrt{3}. \tag{2}$$

We use a codebook $\mathcal{S}$ which is a finite subset of $\mathbb{R}^n$. The receiver decodes the received vector $\mathbf{r}$ to a codeword which optimizes a certain criterion. Two well-known criteria are based on the (squared) Euclidean distance and the Pearson distance.

The classical squared Euclidean distance between the received vector $\mathbf{r}$ and a codeword $\hat{\mathbf{x}} \in \mathcal{S}$ is defined as

$$\delta_E(\mathbf{r}, \hat{\mathbf{x}}) = \sum_{i=1}^{n} (r_i - \hat{x}_i)^2. \tag{3}$$

A Euclidean decoder chooses a codeword minimizing this distance. It is known to be optimal with regard to handling Gaussian noise.

The Pearson distance measure is used in situations which require resistance towards both offset and/or gain mismatch. For any vector $\mathbf{u} \in \mathbb{R}^n$, let $\bar{\mathbf{u}} = (1/n)\sum_{i=1}^{n} u_i$ denote the average symbol value, and let $\sigma_{\mathbf{u}} = (\sum_{i=1}^{n} (u_i - \bar{\mathbf{u}})^2)^{1/2}$ denote the unnormalized symbol standard deviation. The Pearson distance between the received vector $\mathbf{r}$ and codewords $\hat{\mathbf{x}} \in \mathcal{S}$ is defined as

$$\delta_P(\mathbf{r}, \hat{\mathbf{x}}) = 1 - \rho_{\mathbf{r},\hat{\mathbf{x}}} = 1 - \frac{\sum_{i=1}^{n} (r_i - \bar{\mathbf{r}})(\hat{x}_i - \bar{\hat{\mathbf{x}}})}{\sigma_{\mathbf{r}}\sigma_{\hat{\mathbf{x}}}}, \tag{4}$$

where $\rho_{\mathbf{r},\hat{\mathbf{x}}}$ is the well-known Pearson correlation coefficient. A Pearson decoder chooses a codeword minimizing this distance. As shown in [3], when there is no gain mismatch, i.e., $a = 1$, a modified Pearson distance criterion is obtained by removing the division by $\sigma_{\hat{\mathbf{x}}}$ and the irrelevant components $\bar{\mathbf{r}}$ and $\sigma_{\mathbf{r}}$ in the optimization process, i.e.,

$$\delta'_P(\mathbf{r}, \hat{\mathbf{x}}) = \sum_{i=1}^{n} (r_i - \hat{x}_i + \bar{\hat{\mathbf{x}}})^2. \tag{5}$$

3 Maximum Intersection Distance Decoding

In this section, we present a ML decoding criterion for channels with uniform noise and offset mismatch. For a codeword $\hat{\mathbf{x}} = (\hat{x}_1, ..., \hat{x}_n) \in \mathcal{S}$, we define its noise environment

$$U_{\hat{\mathbf{x}}} = \{\mathbf{y} = (y_1, ..., y_n) \in \mathbb{R}^n : \hat{x}_i - \sigma\sqrt{3} < y_i < \hat{x}_i + \sigma\sqrt{3}\}.$$

For a vector $\mathbf{r} \in \mathbb{R}^n$, we write

$$L_{\mathbf{r}} = \{\mathbf{r} - b\mathbf{1} : b \in (-\beta\sqrt{3}, \beta\sqrt{3})\}$$

for the line segment centered at $\mathbf{r}$ with length $2\beta\sqrt{3}$ and direction $\mathbf{1}$.

In order to achieve ML decoding, we need to choose the codeword of maximum probability given the received vector. Assuming all codewords are equally likely, this is equivalent to maximizing the probability density value of the received vector $\mathbf{r}$ given the candidate codeword $\hat{\mathbf{x}}$. From the channel model it easily follows that we should thus maximize

$$\int_{-\beta\sqrt{3}}^{\beta\sqrt{3}} \phi(\mathbf{r} - \hat{\mathbf{x}} - b\mathbf{1})\zeta(b)db. \tag{6}$$

Because of the uniform nature of both ϕ and ζ, this is tantamount to choosing a codeword $\hat{\mathbf{x}}$ for which the noise environment $U_{\hat{\mathbf{x}}}$ has the largest intersection with the line segment $L_{\mathbf{r}}$. Therefore, define the intersection distance $\mathrm{ISD}(\mathbf{r}, \hat{\mathbf{x}})$ between $\mathbf{r}$ and $\hat{\mathbf{x}}$ as the length of the intersection between the noise environment $U_{\hat{\mathbf{x}}}$ and the line segment $L_{\mathbf{r}}$. The most likely candidate codeword $\mathbf{x}_o$ for a received vector has the largest intersection distance, that is

$$\mathbf{x}_o = \arg\max_{\hat{\mathbf{x}} \in S} \mathrm{ISD}(\mathbf{r}, \hat{\mathbf{x}}). \tag{7}$$

Note that a point $\mathbf{r} - t\mathbf{1}$ of $L_{\mathbf{r}}$ is in $U_{\hat{\mathbf{x}}}$ if and only if t satisfies

$$\begin{cases} r_i - \hat{x}_i - \sigma\sqrt{3} & < t < \quad r_i - \hat{x}_i + \sigma\sqrt{3}, \forall i = 1, ..., n, \\ -\beta\sqrt{3} & < t < \quad \beta\sqrt{3}. \end{cases} \tag{8}$$

Defining

$$\begin{aligned} t_0(\mathbf{r}, \hat{\mathbf{x}}) &= \min\left(\{r_i - \hat{x}_i + \sigma\sqrt{3} \,|i = 1, ..., n\} \cup \{\beta\sqrt{3}\}\right), \\ t_1(\mathbf{r}, \hat{\mathbf{x}}) &= \max\left(\{r_i - \hat{x}_i - \sigma\sqrt{3} \,|i = 1, ..., n\} \cup \{-\beta\sqrt{3}\}\right), \end{aligned} \tag{9}$$

we can express the intersection distance between $\mathbf{r}$ and $\hat{\mathbf{x}}$ as

$$\mathrm{ISD}(\mathbf{r}, \hat{\mathbf{x}}) = \sqrt{n(\max\{t_0(\mathbf{r}, \hat{\mathbf{x}}) - t_1(\mathbf{r}, \hat{\mathbf{x}}), 0\})^2}. \tag{10}$$

Note that maximizing $\mathrm{ISD}(\mathbf{r}, \hat{\mathbf{x}})$ is equivalent to maximizing the simplified measure

$$\mathrm{ISD}'(\mathbf{r}, \hat{\mathbf{x}}) = \max\{t_0(\mathbf{r}, \hat{\mathbf{x}}) - t_1(\mathbf{r}, \hat{\mathbf{x}}), 0\}. \tag{11}$$

In Figure 1, we give simulation results for WER of the code

$$\mathcal{S}_1 = \{(0, 0, 0), (1, 1, 0), (1, 0, 1), (0, 1, 1)\}$$

of length $n = 3$ and size 4, with different detectors and various values of the offset standard deviation β, while fixing the noise standard deviation $\sigma = 0.15$. Note that the Pearson decoder has stable performances because of its inherent resistance to offset mismatch, while performances of the Euclidean decoder get worse for increasing values of β. Further, note that in case neither the noise nor the offset is strongly dominating the other, the intersection distance based decoder is clearly outperforming both the Euclidean decoder and the Pearson decoder. For small values of β, the performance curves of the intersection decoder and the Euclidean decoder are gone, which suggests zero WER. This will be further investigated in the next section.

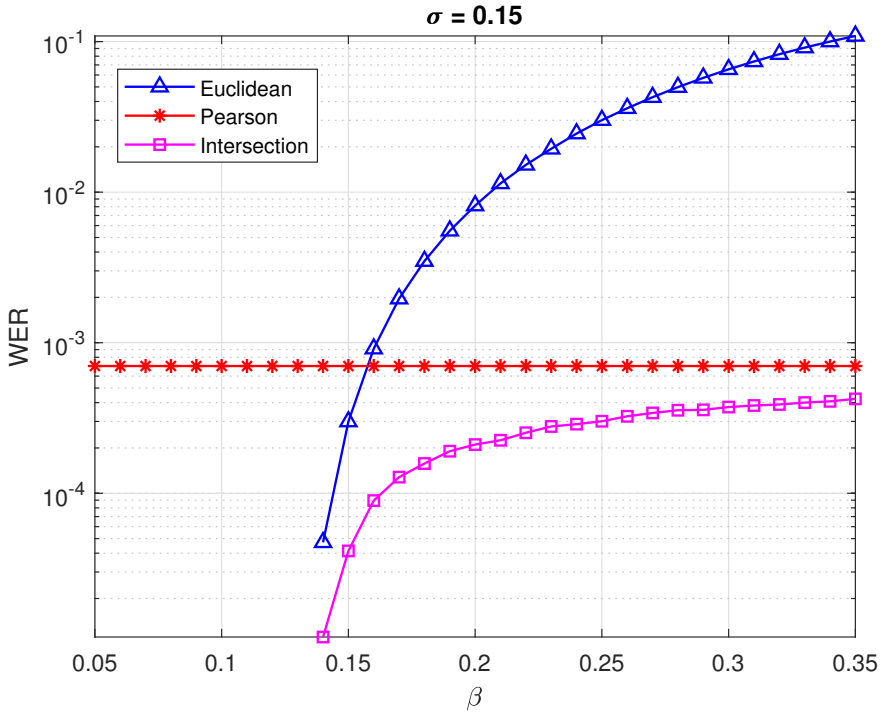


Figure 1: Simulated WER of Euclidean decoding, Pearson decoding, and intersection decoding for the code $\mathcal{S}_1$ without gain mismatch ($a = 1$), uniform offset mismatch with standard deviation β, and uniform noise with standard deviation $\sigma = 0.15$.

4 Zero WER Analysis

Since the noise and the offset are both assumed to be uniformly distributed, it is clear that a WER of zero will be achieved if σ and β are sufficiently small. In this section we present bounds on σ and β guaranteeing zero WER for Euclidean, Pearson, and intersection decoders.

4.1 Euclidean Decoder

When the sum of the noise and offset standard deviations is sufficiently small, the Euclidean decoder can achieve zero WER performance, as shown in the following result.

Theorem 1. *If*

$$\sigma + \beta \leq \min_{\mathbf{s},\mathbf{c}\in\mathcal{S},\mathbf{s}\neq\mathbf{c}} \left(\frac{\sum_{i=1}^{n} (s_i - c_i)^2}{2\sqrt{3} \sum_{i=1}^{n} |s_i - c_i|} \right), \tag{12}$$

then the Euclidean decoder achieves a WER equal to zero.

Proof. Assume that $\mathbf{x} \in \mathcal{S}$ is sent and $\mathbf{r} = \mathbf{x} + \mathbf{v} + b\mathbf{1}$ is received. Then, for all

codewords $\hat{\mathbf{x}} \neq \mathbf{x}$, it holds that

$$\begin{aligned}
&\delta_E(\mathbf{r}, \hat{\mathbf{x}}) - \delta_E(\mathbf{r}, \mathbf{x}) \\
&= \sum_{i=1}^{n} (r_i - \hat{x}_i)^2 - \sum_{i=1}^{n} (r_i - x_i)^2 \\
&= \sum_{i=1}^{n} (r_i - x_i - \hat{x}_i + x_i)^2 - \sum_{i=1}^{n} (r_i - x_i)^2 \\
&= \sum_{i=1}^{n} (\hat{x}_i - x_i)^2 - 2 \sum_{i=1}^{n} (\hat{x}_i - x_i)(r_i - x_i) \\
&= \sum_{i=1}^{n} (\hat{x}_i - x_i)^2 - 2 \sum_{i=1}^{n} (\hat{x}_i - x_i)(v_i + b) \\
&\geq 2(\sigma + \beta)\sqrt{3} \sum_{i=1}^{n} |\hat{x}_i - x_i| - 2 \sum_{i=1}^{n} |\hat{x}_i - x_i| \, |v_i + b| \\
&= 2 \sum_{i=1}^{n} |\hat{x}_i - x_i| \, (\sigma\sqrt{3} + \beta\sqrt{3} - |v_i + b|) \\
&> 0,
\end{aligned}$$

where the first inequality follows from (12) and the last inequality from the fact that $|v_i + b| \leq |v_i| + |b| < \sigma\sqrt{3} + \beta\sqrt{3}$ for all i. Hence, if decoding is based on minimizing (3), the transmitted codeword is always chosen as the decoding result, leading to a WER equal to zero. □

4.2 Pearson Decoder

Since Pearson distance decoding features its immunity to offset mismatch, zero WER performance only requires a limited value of σ, as shown in the next theorem.

Theorem 2. *If*

$$\sigma < \min_{\mathbf{s},\mathbf{c} \in \mathcal{S}, \mathbf{s} \neq \mathbf{c}} \left(\frac{\sum_{i=1}^{n} (s_i - \overline{\mathbf{s}} - c_i + \overline{\mathbf{c}})^2}{\frac{n-1}{n} 4\sqrt{3} \sum_{i=1}^{n} |s_i - \overline{\mathbf{s}} - c_i + \overline{\mathbf{c}}|} \right), \tag{13}$$

then the Pearson decoder achieves a WER equal to zero.

Proof. Assume that $\mathbf{x} \in \mathcal{S}$ is sent and $\mathbf{r} = \mathbf{x} + \mathbf{v} + b\mathbf{1}$ is received. Then, for all

codewords $\hat{\mathbf{x}} \neq \mathbf{x}$, it holds that

$$\begin{aligned}
&\delta'_P(\mathbf{r}, \hat{\mathbf{x}}) - \delta'_P(\mathbf{r}, \mathbf{x}) \\
&= \sum_{i=1}^{n} (r_i - \hat{x}_i + \bar{\hat{\mathbf{x}}})^2 - \sum_{i=1}^{n} (r_i - x_i + \bar{\mathbf{x}})^2 \\
&= \sum_{i=1}^{n} (r_i - \hat{x}_i + \bar{\hat{\mathbf{x}}} - \bar{\mathbf{r}})^2 - \sum_{i=1}^{n} (r_i - x_i + \bar{\mathbf{x}} - \bar{\mathbf{r}})^2 \\
&= \sum_{i=1}^{n} (r_i - x_i + \bar{\mathbf{x}} - \bar{\mathbf{r}} + x_i - \bar{\mathbf{x}} - \hat{x}_i + \bar{\hat{\mathbf{x}}})^2 - \sum_{i=1}^{n} (r_i - x_i + \bar{\mathbf{x}} - \bar{\mathbf{r}})^2 \\
&= \sum_{i=1}^{n} (x_i - \bar{\mathbf{x}} - \hat{x}_i + \bar{\hat{\mathbf{x}}})^2 + 2 \sum_{i=1}^{n} (x_i - \bar{\mathbf{x}} - \hat{x}_i + \bar{\hat{\mathbf{x}}})(r_i - x_i + \bar{\mathbf{x}} - \bar{\mathbf{r}}) \\
&= \sum_{i=1}^{n} (x_i - \bar{\mathbf{x}} - \hat{x}_i + \bar{\hat{\mathbf{x}}})^2 + 2 \sum_{i=1}^{n} (x_i - \bar{\mathbf{x}} - \hat{x}_i + \bar{\hat{\mathbf{x}}})(x_i + v_i + b - x_i + \bar{\mathbf{x}} - \bar{\mathbf{x}} - \bar{\mathbf{v}} - b) \\
&= \sum_{i=1}^{n} (x_i - \bar{\mathbf{x}} - \hat{x}_i + \bar{\hat{\mathbf{x}}})^2 + 2 \sum_{i=1}^{n} (x_i - \bar{\mathbf{x}} - \hat{x}_i + \bar{\hat{\mathbf{x}}})(v_i - \bar{\mathbf{v}}) \\
&> \tfrac{n-1}{n} 4\sigma\sqrt{3} \sum_{i=1}^{n} \left| x_i - \bar{\mathbf{x}} - \hat{x}_i + \bar{\hat{\mathbf{x}}} \right| - 2 \sum_{i=1}^{n} \left| x_i - \bar{\mathbf{x}} - \hat{x}_i + \bar{\hat{\mathbf{x}}} \right| |v_i - \bar{\mathbf{v}}| \\
&= 2 \sum_{i=1}^{n} \left| x_i - \bar{\mathbf{x}} - \hat{x}_i + \bar{\hat{\mathbf{x}}} \right| (\tfrac{n-1}{n} 2\sigma\sqrt{3} - |v_i - \bar{\mathbf{v}}|) \\
&\geq 0.
\end{aligned}$$

where the first inequality follows from (13) and the last inequality from the fact that $|v_i - \bar{\mathbf{v}}| < \frac{n-1}{n} 2\sigma\sqrt{3}$ for all i. Hence, if decoding is based on minimizing (5), the transmitted codeword is always chosen as the decoding result, leading to a WER equal to zero. □

4.3 Intersection Decoder

In this subsection, we show that zero WER is achieved for the intersection decoder in the case that σ or $\sigma + \beta$ is sufficiently small.

Theorem 3. *If*

$$\sigma \leq \min_{\mathbf{s},\mathbf{c} \in \mathcal{S}, \mathbf{s} \neq \mathbf{c}} \left(\frac{\max\limits_{1 \leq i,j \leq n} \{(s_i - c_i) - (s_j - c_j)\}}{4\sqrt{3}} \right) \tag{14}$$

or

$$\sigma + \beta \leq \min_{\mathbf{s},\mathbf{c} \in \mathcal{S}, \mathbf{s} \neq \mathbf{c}} \left(\frac{\max\limits_{i=1,\ldots,n} (|s_i - c_i|)}{2\sqrt{3}} \right) \tag{15}$$

then the intersection decoder achieves a WER equal to zero.

Proof. Assume that $\mathbf{x} \in \mathcal{S}$ is sent and $\mathbf{r} = \mathbf{x} + \mathbf{v} + b\mathbf{1}$ is received. We will show that if (14) or (15) holds, then $\mathrm{ISD}'(\mathbf{r}, \hat{\mathbf{x}}) = 0$ for all codewords $\hat{\mathbf{x}} \neq \mathbf{x}$. First of all, note that

$$\begin{aligned}
&t_0(\mathbf{r}, \hat{\mathbf{x}}) - t_1(\mathbf{r}, \hat{\mathbf{x}}) \\
&= \min\left(\{r_i - \hat{x}_i + \sigma\sqrt{3} \,|i = 1, \ldots, n\} \cup \{\beta\sqrt{3}\}\right) \\
&\qquad - \max\left(\{r_i - \hat{x}_i - \sigma\sqrt{3} \,|i = 1, \ldots, n\} \cup \{-\beta\sqrt{3}\}\right) \\
&= \min\left(\{r_i - \hat{x}_i + \sigma\sqrt{3} \,|i = 1, \ldots, n\} \cup \{\beta\sqrt{3}\}\right) \\
&\qquad + \min\left(\{-(r_i - \hat{x}_i) + \sigma\sqrt{3} \,|i = 1, \ldots, n\} \cup \{\beta\sqrt{3}\}\right) \\
&= \min(\{2\beta\sqrt{3}\} \cup \{\min_{i=1,\ldots,n} \{-|r_i - \hat{x}_i|\} + \sigma\sqrt{3} + \beta\sqrt{3}\} \\
&\qquad \cup \{\min_{1 \leq i,j \leq n} \{(r_i - \hat{x}_i) - (r_j - \hat{x}_j)\} + 2\sigma\sqrt{3}\}).
\end{aligned} \tag{16}$$

Next, we will show that if (14) or (15) holds, this expression is negative whenever $\hat{\mathbf{x}} \neq \mathbf{x}$.

If (14) holds, then

$$\begin{aligned}
&\min_{1\leq i,j\leq n}\{(r_i-\hat{x}_i)-(r_j-\hat{x}_j)\}+2\sigma\sqrt{3}\\
&=\min_{1\leq i,j\leq n}\{(r_i-\hat{x}_i)-(r_j-\hat{x}_j)\}-2\sigma\sqrt{3}+4\sigma\sqrt{3}\\
&<\min_{1\leq i,j\leq n}\{(r_i-\hat{x}_i)-(r_j-\hat{x}_j)-(v_i-v_j)\}+4\sigma\sqrt{3}\\
&=\min_{1\leq i,j\leq n}\{[(r_i-\hat{x}_i)-(r_j-\hat{x}_j)]-[(r_i-x_i-b)-(r_j-x_j-b)]\}+4\sigma\sqrt{3}\\
&=\min_{1\leq i,j\leq n}\{(x_i-\hat{x}_i)-(x_j-\hat{x}_j)\}+4\sigma\sqrt{3}\\
&=-\max_{1\leq i,j\leq n}\{(\hat{x}_i-x_i)-(\hat{x}_j-x_j)\}+4\sigma\sqrt{3}\\
&\leq 0.
\end{aligned} \tag{17}$$

where the first inequality follows from the fact that $v_i - v_j \leq |v_i| + |v_j| < 2\sigma\sqrt{3}$ and the second inequality from (14).

If (15) holds, then

$$\begin{aligned}
&\min_{i=1,\ldots,n}\{-|r_i-\hat{x}_i|\}+\sigma\sqrt{3}+\beta\sqrt{3}\\
&=\min_{i=1,\ldots,n}\{-|r_i-\hat{x}_i|\}-\sigma\sqrt{3}-\beta\sqrt{3}+2\sqrt{3}(\sigma+\beta)\\
&<\min_{i=1,\ldots,n}\{-|r_i-\hat{x}_i|-|v_i+b|\}+2\sqrt{3}(\sigma+\beta)\\
&=\min_{i=1,\ldots,n}\{-|r_i-\hat{x}_i|-|r_i-x_i|\}+2\sqrt{3}(\sigma+\beta)\\
&\leq\min_{i=1,\ldots,n}\{-|x_i-\hat{x}_i|\}+2\sqrt{3}(\sigma+\beta)\\
&=-\max_{i=1,\ldots,n}\{|x_i-\hat{x}_i|\}+2\sqrt{3}(\sigma+\beta)\\
&\leq 0
\end{aligned} \tag{18}$$

where the first inequality follows from the fact that $|v_i + b| \leq |v_i| + |b| < \sigma\sqrt{3} + \beta\sqrt{3}$ and the last inequality from (15).

Combining (11), (16), (17), and (18), we find that indeed $\mathrm{ISD}'(\mathbf{r}, \hat{\mathbf{x}}) = 0$ for all codewords $\hat{\mathbf{x}} \neq \mathbf{x}$. By definition, $\mathrm{ISD}'(\mathbf{r}, \mathbf{x}) > 0$. This implies that if decoding is based on maximizing (11), the transmitted codeword is always chosen as the decoding result, leading to a WER equal to zero. □

Simulated WER with various values of σ and β for the example code $\mathcal{S}_1$ from the previous section are shown in Fig.2. On one hand, we find for $\beta = 0.2$ and $\beta = 0.15$ that zero WER is achieved when $\sigma < 1/(4\sqrt{3}) = 0.144$, which agrees with the bound from (14). On the other hand, we find for $\beta = 0.1$ and $\beta = 0.05$ that zero WER is achieved when $\sigma + \beta < 1/(2\sqrt{3}) = 0.288$, which agrees with the bound from (15).

5 Conclusion

In this paper, maximum likelihood decoding for channels with uniform noise and offset mismatch has been presented. This method, which is based on the intersection distance, combines the immunity to offset mismatch of Pearson distance decoding and the higher noise resistance of Euclidean distance decoding. It has been shown that

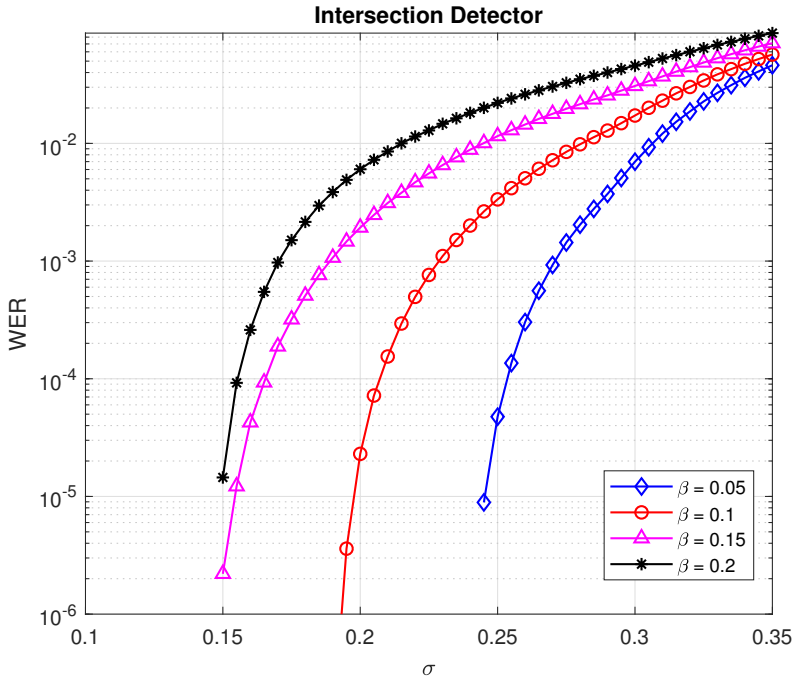


Figure 2: Simulated WER of intersection decoding for the code $\mathcal{S}_1$ in the case no gain mismatch ($a = 1$), uniform offset mismatch with standard deviation β, and uniform noise with standard deviation σ.

for sufficiently small standard deviations of the noise and/or offset zero WER can be achieved.

For future work, we are interested in investigating maximum likelihood decoding for noise channels with known distribution of both offset and gain mismatch.

References

[1] G. Bouwhuis, J. Braat, A. Huijser, J. Pasman, G. van Rosmalen, and K. A. S. Immink, *Principles of Optical Disc Systems*. Boston, MA, USA: Adam Hilger, 1985.

[2] A. Jiang, R. Mateescu, M. Schwartz, and J. Bruck, "Rank modulation for flash memories," IEEE Transactions on Information Theory, vol. 55, no. 6, pp. 2659–2673, June 2009.

[3] K. A. S. Immink and J. H. Weber, "Minimum Pearson Distance Detection for Multilevel Channels with Gain and/or Offset Mismatch," IEEE Transactions on Information Theory, vol. 60, no. 10, pp. 5966–5974, Oct. 2014.

[4] S. R. Blackburn, "Maximum Likelihood Decoding for Multilevel Channels with Gain and Offset Mismatch," IEEE Transactions on Information Theory, vol. 62, no. 3, pp. 1144–1149, March 2016.

[5] J. H. Weber and K. A. S. Immink, "Maximum Likelihood Decoding for Gaussian Noise Channels with Gain or Offset Mismatch," to appear in IEEE Communications Letters, 2018.

Measurement-based Assessment of Noise Sources in Office and Household Environments Impacting Ultrasound Indoor Positioning

Chesney Buyle Bert Cox Liesbet Van der Perre
KU Leuven
ESAT, DRAMCO
Gebroeders De Smetstraat 1, BE-9000 Ghent
name.surname@kuleuven.be

Abstract

Interference due to ambient noise is an important aspect in precise and accurate, ultrasonic distance measurements. In indoor positioning, interference can modify or disrupt a signal that is travelling along a channel between source and receiver, resulting in false or even the absence of a distance measurement. This paper introduces an ultrasonic recording circuit based on Microelectromechanical Systems (MEMS), to analyze whether it is possible to bypass the potential ambient ultrasonic noise sources in office and household environments. Measurements with this system show the sonic spectrum up to 200 kHz of several everyday objects. The results of these measurements are of great importance towards future exploration of ultrasonic positioning.

1 Introduction

Some positioning systems have the purpose to determine a legitimite area in which the target is located. Temperature, humidity [1] and interference are the three main causes having a negative influence on the precision and accuracy in ultrasonic, and more generally sound based, distance measurements. For the first two, formulas can be integrated in the system to cope with this problem. The latter however can modify or disrupt a signal that is travelling along a channel between source and receiver, resulting in false or even the absence of a distance measurement. Moreover, ultrasonic noise causing these interferences are highly dependant of the ambient surrounding. It is possible to distinguish two types of ultrasonic noise sources [2]: technological ultrasonic noise sources use ultrasound to execute or improve certain processes, e.g. dust removal, aerosol production, physical therapy, etc. The second type are sources that unintentionally generate ultrasound, better known as non-technological noise sources.

Previous research mainly focused on the potential hazards of the exposure to ultrasound on our auditory system [3]. This research was oriented towards industrial, e.g. ultrasonic welding [4], and health applications e.g. ultrasonic scaler [5]. Measurements were performed in these environments. Another class of ultrasonic ambient noise measurements can be found in bio-engineering, where the ultrasonic noises created by insects in rain forests are analyzed [6] and a characterization of ultrasonic ambient sea noise in shallow water can be found in [7]. Even NASA is doing Ultrasonic Background Noise Tests (UBNT) on the ISS to identify and locate high pitched sounds created by air leaking through a pressurized wall.
However, to our knowledge, no research was yet performed on the ambient ultrasonic noise sources in office and household environments, both important settings for the deployment of indoor navigation systems. To analyze whether it is possible to bypass

the potential interference in these surroundings, an ultrasonic recording circuit was created based on Microelectromechanical Systems (MEMS) and is explained in section 2. The next section proposes the measurements performed with this system and gain insight in the several ultrasonic noise sources. A conclusion and potential, future work are given in the last section.
This research was the initial step towards a low-power hybrid signaling location verification system. The ambient measurements described here, were performed to see whether the broadcasted, ultrasonic signal is not affected by surrounding noise.

2 MEMS-based, Mobile, US Recording System

2.1 MEMS Microphone

One of the main assets in ultra-low power acoustic positioning systems, in comparison to the RF-based ones, is the improved accuracy obtained at the lower system frequency due to the lower propagation speed of mechanical waves. Moreover, since the introduction of *Microelectromechanical Systems* (MEMS), modules can nowadays be realized at low cost with a small form factor and can compeed with other distance measurement systems on a sensor energy level. The sensor chosen in this paper that converts the ambient sound signals into electrical signals is the Knowles SPU0410HRH5H-PB MEMS microphone [8]. This unique MEMS microphone has a frequency response of over the audible threshold of 20 kHz.
An important specification when working with microphones is the sensitivity. It measures the size of the output signal at a certain Sound Pressure Level (SPL). In section 2.2, the received sound signals are amplified and filtered. Therefore, it is more convenient in this paper to work with these voltages. Using equation (1) it is possible to convert the given dBV/Pa into mV/Pa.

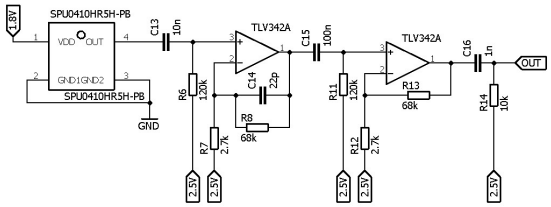
$$Sensitivity\,[dBV] = 20\log\left(\frac{Sensitivity\,[mV/Pa]}{1\,V/Pa}\right) \tag{1}$$

The SiSonic microphone has a sensitivity between $-45\,\mathrm{dBV/Pa}$ (=5.62 mV/Pa) and $-39\,\mathrm{dBV/Pa}$ (=11.22 mV/Pa) at a sinusoidal input signal of 1 kHz at 94 dB SPL. In comparison, a lawn moyer or a busy street have a SPL of 90 dB.

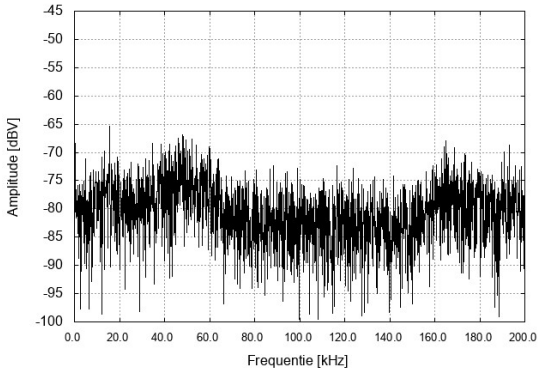
2.2 Amplification and filtering

For further optimized processing, the output signals should be in the range of Volts instead of millivolts. Amplification is thus necessary. Figure 1a shows the amplification and filtering circuit. A large unity gain bandwidth, imposed by the high frequent, low voltage output signals, in combination with low power consumption can be found in the TLV342A amplifier. The input and feedback resistor at the TLV342 are chosen for a bandwidth of 80 kHz. The closed loop amplification in these cases is 26.18. As a single amplifier is not enough, two non-inverting operational amplifiers are cascaded. The maximum amplification in this scenario is 56.72 dB. Biasing the audio output voltage is necessary for maximal output swing.
For noise reduction, a first order high pass filter is added to the microphone and both amplifier outputs. A spice simulated transfer function can be seen in figure 2. The maximum amplification here is 54.25 dB at 27.45 kHz, conform the theoretical value.

Figure 1



(a) Amplification and filtering circuit.



(b) Reference frequency spectrum with no electrical appliances in the neighbourhood.

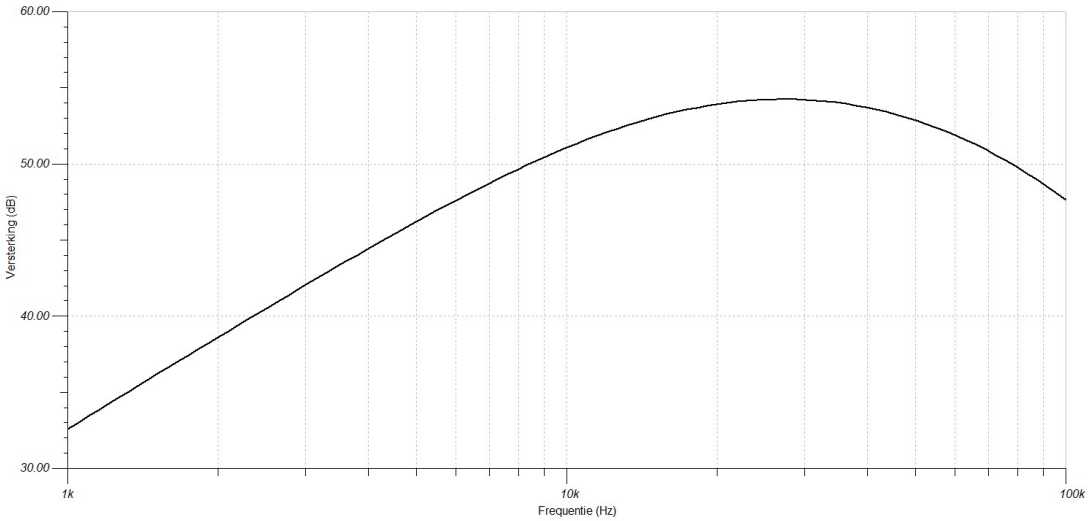


Figure 2: Spice simulated transfer function of the circuit.

3 Measurements

In an ordinary office or household, most electrical equipment uses a lower, DC voltage other than the available mains voltage. To convert this AC voltage to the desired DC voltage, a switched power supply is often used. Coil windings and capacitors in these circuits are a potential source of ultrasonic noise. To obtain the frequency spectrum, the circuit depicted in figure 1a is linked to an Agilent DSO-X 2002 digital oscilloscope on which Fast Fourier Transforms are performed to obtain the frequency response. Table 1 shows the scope settings of the several measurements. The circuit is positioned several centimeters away from the different sources. The spectra of these measurements are depicted in the figures on the following page. A reference spectrum of the room is given in figure 1b.

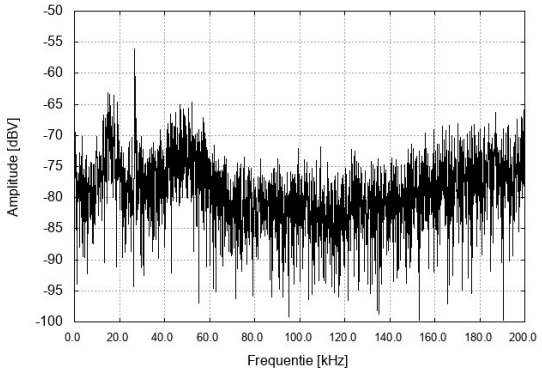
Table 1: Scope settings of Agilent DSO-X 2002.

Volts/div	**Time/div**	**Scale**	**Δf**	**Window**
1 V/div	5 ms/div	10dB	0.1 kHz	rectangle

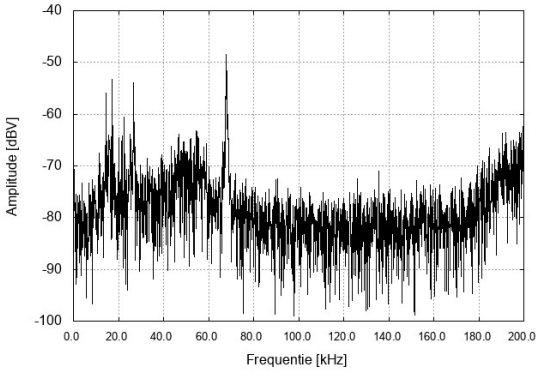
Desktop Computer: Dell Optiflex 9020

Two types of measurements are performed on the desktop computer. The first one is where the PC is powered off, but the supply remains connected. In this case, a peak can be detected at 26.5 kHz. Subsequently the PC is powered on. Next to the earlier mentioned maximum, an additional larger peak can be identified at 67.5 kHz and some smaller ones just above the 20 kHz. To analyse the influence of the distance on the spectrum, the circuit was moved to 50 cm and 1 m (Fig. 3c and 3d respectively). In both cases, the frequency component at 26.5 kHz disappeared.

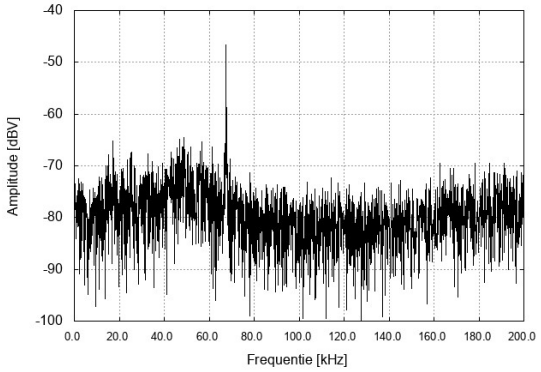
Figure 3



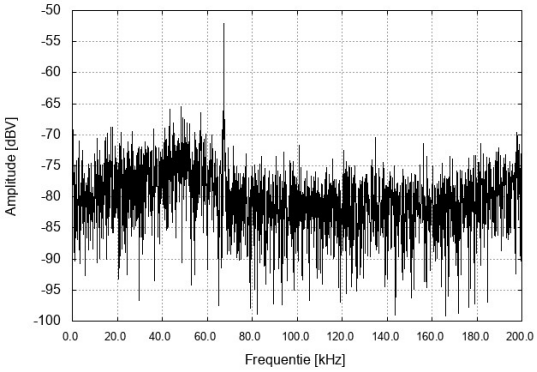
(a) Frequency spectrum of PC when turned off, plugged in to the mains power.



(b) Frequency spectrum of PC when turned on, 5 cm away from the backside.



(c) Frequency spectrum of PC when turned on, 50 cm away from the backside.

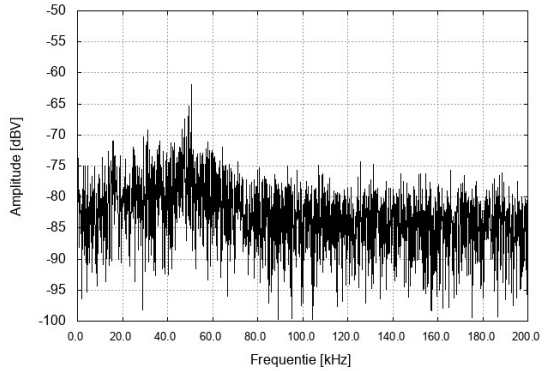


(d) Frequency spectrum of PC when turned on, 1 m away from the backside.

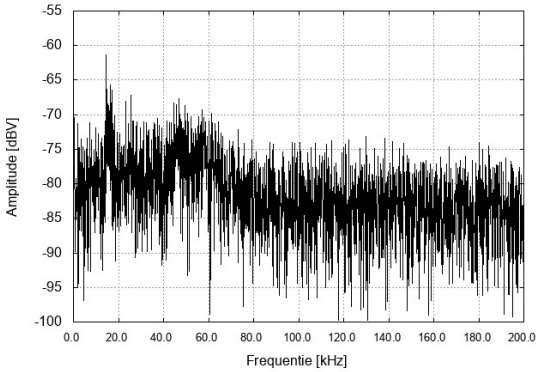
Laptop: Acer Aspire V17 Nitro and Power Supply

If we compare the spectrum of the laptop in stand-by mode to the reference spectrum, no dominant frequency can be distinguished. However, when the laptop is turned on, a maximum at 50 kHz can be observed when the circuit is placed in front of it. The response of the accompanied power supply with the laptop in stand-by and power-on mode can be found in figure 4c and 4c respectively. In the first case, the frequency components can be found around 35.5 kHz and 40 kHz, and around 26.5 kHz and in the 40.4 kHz - 44.6 kHz in the second case.

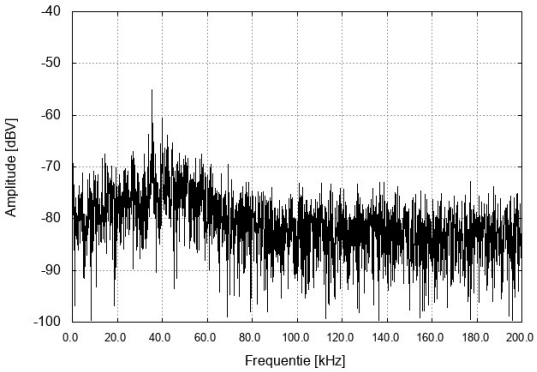
Figure 4



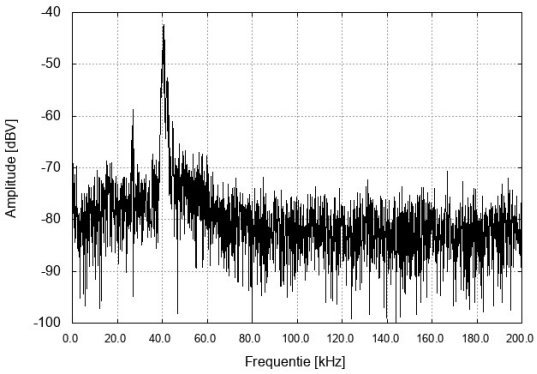
(a) Frequency spectrum of the laptop in stand-by modus.



(b) Frequency spectrum at the screen of the laptop when turned on.



(c) Frequency spectrum of the laptop power supply, when the laptop is in stand-by modus.

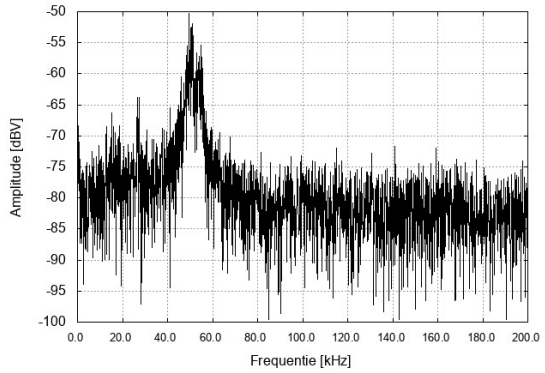


(d) Frequency spectrum of the laptop power supply, when the laptop is powered on.

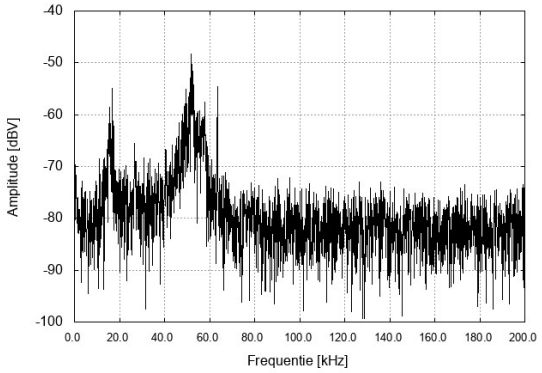
Television: Telefunken L48F249X3CW-3DU

Another item that can be found in every household is the television. The Telefunken type used in these measurements is a LED TV. Again, the spectrum is measured in stand-by and power-on mode. Comparing the stand-by spectrum with the reference spectrum does not show large differences. In this situation, no peaks are noticed. In power-on, maximum values can be detected at 26.5 kHz and in the interval of 45.9 kHz to 56.6 kHz when we measure in front of the television. At the back of the TV, an additional peak is observed at 63.3 kHz.

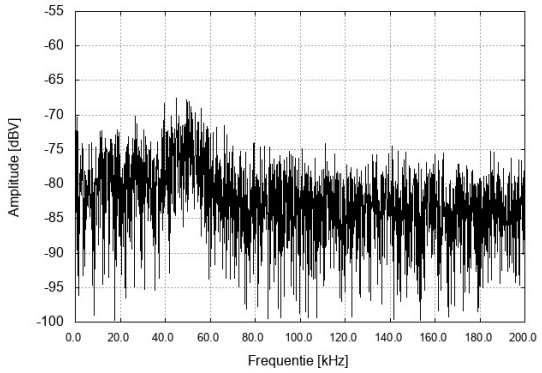
Figure 5



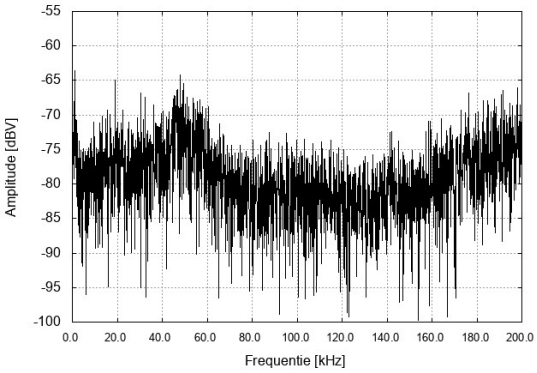
(a) Frequency spectrum of the television on the frontside, when powered on.



(b) Frequency spectrum of the television on the backside, when powered on.



(c) Frequency spectrum of the television in stand-by modus.

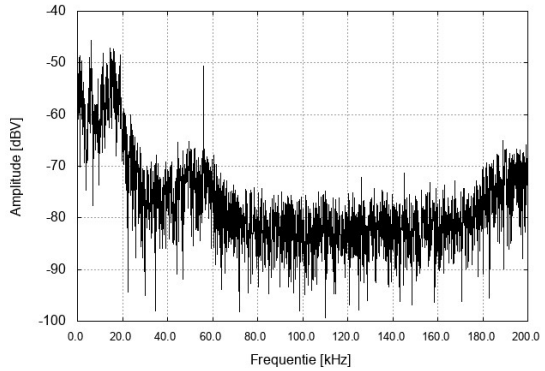


(d) Frequency spectrum of the printer in stand-by modus.

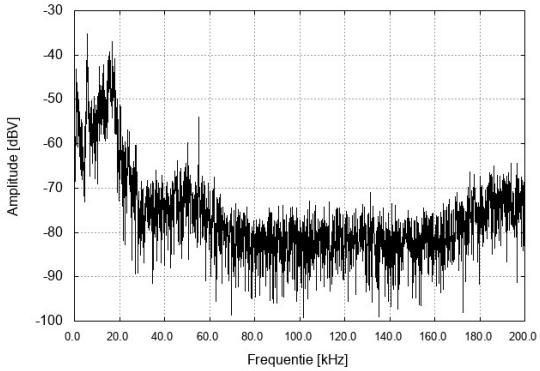
Printer: Brother HL-5350DN

In this household/office element, no real maxima can be detected in stand-by modus as well. In start-up, the time period between switching on the printer and the moment the printer can accept a print job, a sharp peak can be detected at 55.9 kHz and some smaller ones just under the 20 kHz. Similar spectra can be obtained when the distance is increased to 50 cm and 1 m.

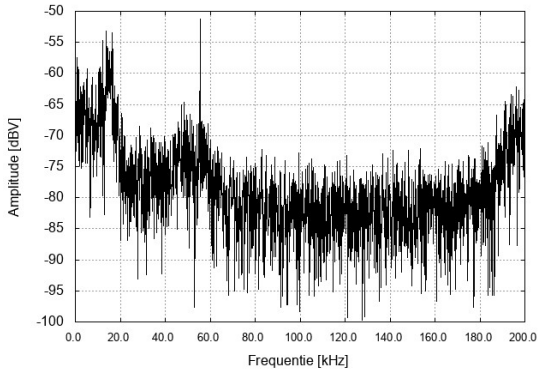
Figure 6



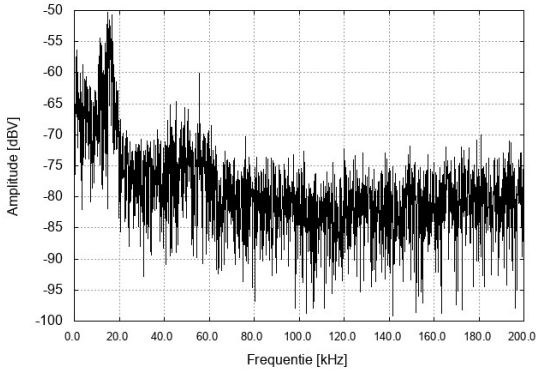
(a) Frequency spectrum of the printer when starting up.



(b) Frequency spectrum of the printer when printing



(c) Frequency spectrum, 50 cm away from the printer, while printing.

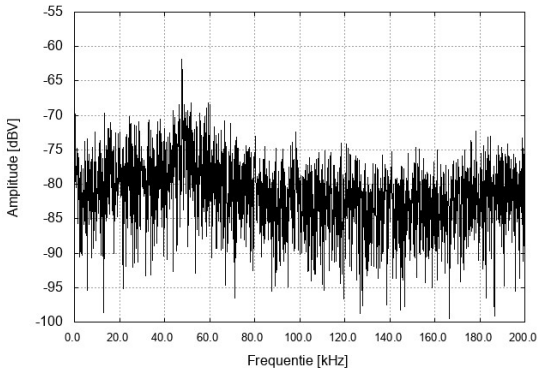


(d) Frequency spectrum, 1 m away from the printer, while printing.

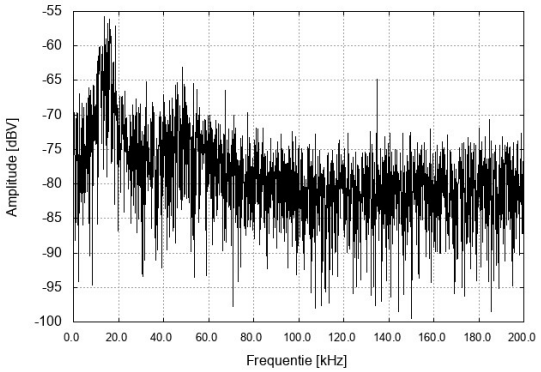
Beamer: Infocus IN3116

A solid peak can be distinguished at 47.7 kHz when the beamer is in stand-by modus (Fig. 7a). In start-up this peak lies around the 134.8 kHz. Important to note is that figure 7b is a snapshot of the spectrum. Other peaks change fast over time and occur over the whole spectrum. The spectrum after start-up has minimal changes towards the reference spectrum.

Figure 7



(a) Frequency spectrum of the beamer when in stand-by modus.

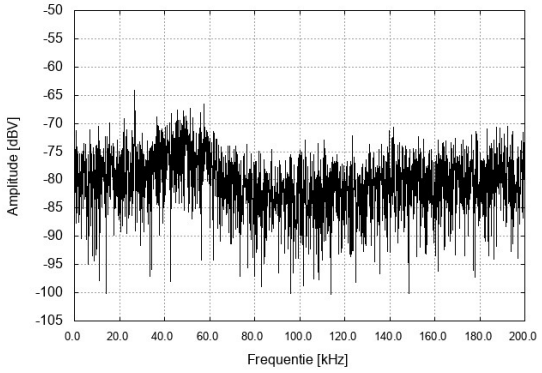


(b) Frequency spectrum of the printer when starting up.

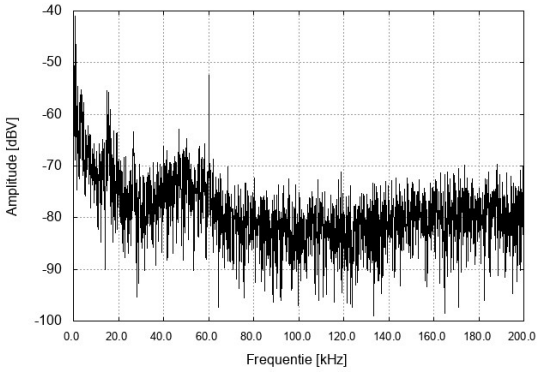
Induction Cooking Plate

The last household item is an induction cooking plate, used in several energy modes. In stand-by, a smaller peak can be detected at 26 kHz. After placing a pan on the plate and turning it in low power mode, a larger peak at 60.0 kHz is observed and noise below the 20 kHz can be heared. This changes drastically when the power level is increased. Periodic peaks with decreasing levels at increased frequency can be distinguished. In the highest energy mode, the levels of the maxima are 30 to 55 dB larger than the previous mentioned spectra. These larger maxima can be detected as well when the cooking pot is taken away in a lower energy mode.

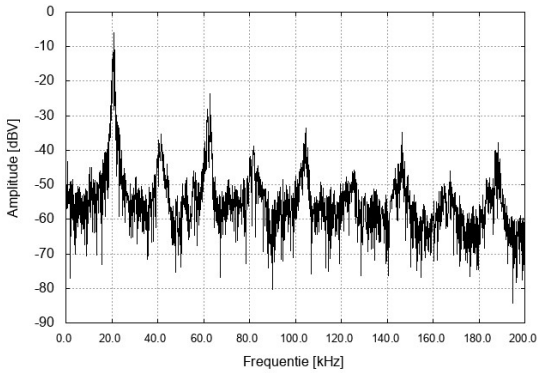
Figure 8



(a) Frequency spectrum of the induction cooking plate in stand-by.



(b) Frequency spectrum of the induction cooking plate in low-energy mode.

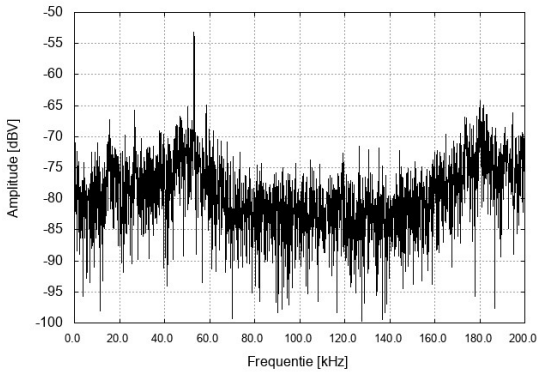


(c) Frequency spectrum of the induction cooking plate in low-energy mode.

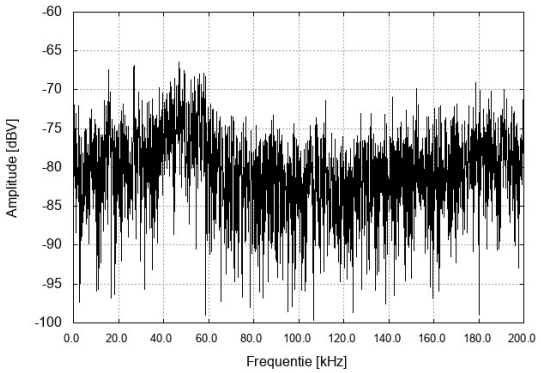
Fluorescent Lamps

The last measurements were performed on fluorescent lamps, commonly used in office buildings. In the initial measurements, the MEMS-circuit is positioned several centimeters away from the lamps. The spectrum in this case has two important values: 26.5 kHz and 53 kHz. Next, the circuit was positioned on 2.5m right under the lamps, to determine the influence of the distance. The earlier mentioned frequencies are still available in the spectrum, at a lower level. Figure 9c shows what happens to the spectrum if the cicrcuit is positioned between two lamps, still on a 2.5 m distance. A sharp peak can still be found at 26.5 kHz.

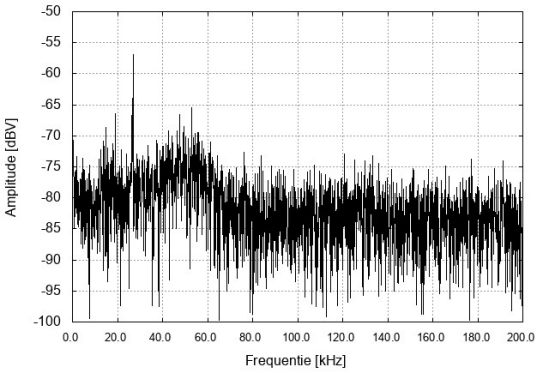
Figure 9



(a) Frequency spectrum just underneath the fluorescent lamps.



(b) Frequency spectrum 2.5 m underneath the fluorescent lamps.



(c) Frequency spectrum 2.5 m underneath and in between fluorescent lamps.

4 Conclusion

It can be concluded that most household and office equipment produce (ultra)sound signals, which can interfere with sound used for ultrasonic positioning. The sound levels of these signals are in, all except one case, lower than -45 dBV. In the case of the induction cooking plate, the measured peaks go up to -8 dBV. However, the only frequency component that is available at several devices is the 26.5 kHz signal. These appliances (television, laptop, desktop computer and induction cooking plate) have a (build-in) power supply generating this sound signal. One can notice that the peaks in these spectra are narrow. Positioning or data transmission errors due to small band signals can be overcome by selecting a different broadcast frequency or by using spread spectrum techniques like DSSS and FHSS.

References

[1] K. G. Panda, D. Agrawal and A. Nshimiyimana, "Effects of environment on accuracy of ultrasonic sensor operates in millimetre range," Phil. Perspectives in Scienceo, pp. 574–576, 2016

[2] B. Smagowska, "Ultrasonic Noise Sources in Work Environment," Archive of Acoustics, vol. 38, pp. 169–176, 2013

[3] Pawlaczyk-Luszczynska M., Dudarewicz A. and Sliwinska-Kowalska M.,"Theoretical Predictions and Actual Hearing Threshold Levels in Workers Exposed to Ultrasonic Noise of Impulsive Character- A Pilot Study," in JOSE, vol. 13, pp. 409418, 2007

[4] A. Dudarewicz, K. Zaborowski, P. Rutkowska-Kaczmarek, M. Zamojska-Daniszewska, M. Sliwinska-Kowalska, A. Zamyslowska-Szmytke and M. Pawlaczyk-Lusczczynska, "The Hearing Threshold of Employees Exposed to Noise Generated by the Low-Frequency Ultrasonic Welding Devices," Archive of Acoustics, vol. 42, pp. 199–206, 2017.

[5] M. Doestzada, "Het geluidsniveau van de ultrasone scaler in de tandheelkunde," Rijksuniversiteit Groningen, 2016

[6] J. S. Saby and H. A. Thorpe,"Ultrasonic Ambient Noise in Tropical Jungles," in The Journal of the Acoustical Soc. of America, Vol 18, pp 271273, 1946

[7] D. H. Cato and M. J. Bell,"Ultrasonic Ambient Noise in Australian Shallow Waters at Frequencies up tp 200kHz," Materials Research Laboratory, Sydney, 1992

[8] Knowles Electronics, "SiSonic SPPU0410HR5H," Product Data Sheet, Itasca, 2013

Long Range IoT Connections: Experimental Confirmation of the Energy Drain and Exploration of Escape Routes

Gilles Callebaut Guus Leenders Geoffrey Ottoy
Lieven De Strycker Liesbet Van der Perre

KU Leuven, ESAT-DRAMCO, Ghent Technology Campus
Ghent, Belgium
gilles.callebaut@kuleuven.be

Abstract

Long range wireless connectivity opens the door for new IoT applications. Low energy consumption is essential to enable long autonomy of devices powered by batteries or devices relying on harvested energy. This paper assesses the discrepancy of satisfying both low power and long range requirements. A versatile module has been developed to assess and optimize power modes in order to maximize the autonomy of IoT devices. A cross-layer approach is put forward to determine the optimal packet lengths for energy efficiency, considering potential collisions and retransmissions. To further increase the energy efficiency of devices in the field, the exploitation of (large) antenna arrays at the side of the Base Station, is explored.

I. Introduction

It is evident from electromagnetic wave propagation physics, that long range and low power make contradictory requirements. Despite this discrepancy, a large subset of the IoT applications require battery-powered devices to be capable of transmitting small messages over large distances. To address the demand, dedicated networks operating in unlicensed bands below 1 GHz, such as LoRaWAN [1] and Sigfox [2], are being rolled out. Also new cellular communication modes and terminal categories are defined, tailored for Machine Type Communication and IoT applications [3], [4].

To adequately validate these technologies a versatile IoT module with Low Power Wide Area Network (LPWAN) connectivity has been developed. It allows researchers to assess and optimize power modes in order to maximize the autonomy of IoT devices, more specifically, for LoRaWAN communication.

LoRaWAN differs from its cellular equivalents by operating in license-exempt radio bands. Furthermore, it focusses on transmitting small messages with a low data rate. In contrast, NB-IoT and LTE-M operate in licensed bands coordinated by the network operators. In addition, they provide a higher data rate and are less constrained in latency and bidirectional communication requirements. Sigfox is more limited in the number of uplink and downlink messages and has a lower data rate compared to LoRaWAN. It has a maximum uplink payload size of 12 Bytes. Sigfox increases its Quality-of-Service by deploying their network in order to receive uplink data by at least three base stations.

The remainder of the paper is structured as follows. First, the developed IoT module facilitating low power development is elaborated. The power consumption and energy profile of the module is assessed and characterized. Secondly, tools designed based on the

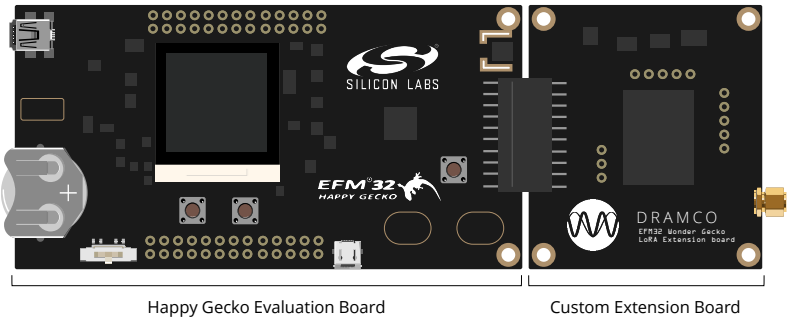


Fig. 1: A Silabs STK3400 "Happy Gecko" together with a custom-design extension board creates a versatile IoT module that supports different sensors and performs real-time power monitoring. Both software and hardware are open-source: github.com/DRAMCO/LoRaWAN_EFM32.

derived IoT node profile are briefly discussed. To conclude, two escape routes to improve the autonomy of energy-constrained LPWAN devices are explored.

II. Versatile IoT module facilitates low power development

In this section the versatile IoT module with LPWAN capabilities, depicted in Figure 1, will be further elaborated upon. This module supports a variety of sensors and performs real-time power monitoring. The module comprises two entities:

1) an EFM32 Happy Gecko developer board hosting a low power computing environment and
2) a custom LoRa extension board providing wireless communication.

The EFM32 Happy Gecko developer board combines the powerful, yet low power, ARM Cortex M0+ with power sensing techniques. This makes the module an ideal development board for use in a very limited power budget system. The custom LoRa based extension board features a LoRa transceiver, i.e. a Semtech SX1272 chip. This radio chip is combined with a variety of sensors, i.e., temperature, humidity, light, movement, tactile, and energy consumption. This combination makes this module a truly versatile module for LPWAN IoT networks. The software and hardware are open-source [5]. The energy consumption can be observed through the IDEs built-in energy profiler.

A. Energy Profiling of an IoT Node reveals the transmitter big spender

The power dominance of wireless communication, and more specifically transmitting data, is illustrated in Figure 2 and Table III. The total energy consumption on the node will determine its battery life. As a consequence, we provide an overall assessment and breakdown of the energy consumption of a typical IoT node in various realistic scenarios. This assessment includes measurements of the energy per bit (Tx) when transmitting messages at different data rates, as depicted in Figure 3. The results clearly indicate the increased energy when using a lower data rate (i.e., more time on air). The time on air dependency on the spreading factor (SF) originates from the defined symbol duration in LoRaWAN:

$$T_{sym} = \frac{2^{\mathrm{SF}}}{\mathrm{BW}} \tag{1}$$

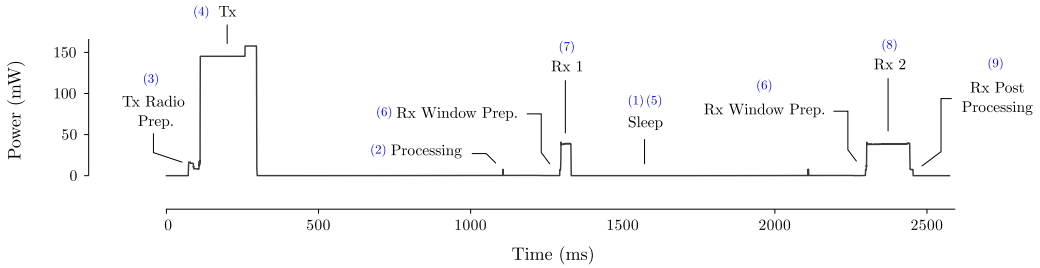


Fig. 2: Measured power states of a LoRa node [5]. These measurements are also summarized in Table II. This profile clearly shows the energy impact of transmitting a message. In this case a confirmed message was sent with SF9 and a payload of 32 bytes.

In addition, Figure 3 also illustrates that accumulating data prior to transmitting can increase the power efficiency of a device. These measurements allow us to accurately predict the battery life of IoT nodes. Furthermore, these results provide a baseline for evaluating our cross-layer optimizations, as discussed in Section III-A.

In addition to decreasing the time on air, other approaches need to be considered when reducing the overall energy consumption of the node. For the sake of completeness, we also briefly discuss additional approaches to efficiently make use of the limited energy-budget.

A first approach is to put all unnecessary components in sleep or even turn them off when not needed. This technique is also applicable to the utilized LoRaWAN modem or chip. This is certainly the case if the duty cycle is very low, i.e. when the node is transmitting with a low periodicity. Providing low duty-cycle transmissions, the power consumption originating from operations between transmissions is no longer negligible and even becomes an important factor.

Another energy-saving approach is to only transmit data when needed. Introducing processing to distinguish relevant from non-relevant information is imperative in power-constrained devices. Based on the relevance or quality, the node can transmit, discard or accumulate the data. Due to its energy-optimized design, the developed node gives a more accurate representation of real IoT nodes, in contrast to existing evaluation boards.

An energy model has been derived based on the conducted measurements. This model can thereafter be utilized in simulations and other tools estimating the energy usage of LPWAN nodes. For instance, a tool[1] has been developed to address the challenge of determining the lifespan of nodes. The lifespan is estimated based on the derived energy consumption model and the capacity delivered by a battery. An indication of the expected battery lifetime of LoRaWAN devices for different IoT applications [6] is shown in Table I. In addition, a simulation framework [7] has been designed, based on this model, in order to also include retransmissions and collisions.

III. Escape Routes to Improve the Autonomy of Energy-Constrained LPWAN Devices

The autonomy LPWAN devices can be improved by:

1) Minimizing the “on air” time. A cross-layer approach is advocated to determine optimal packet lengths for energy efficiency, considering potential collisions and retransmissions.

[1]https://dramco.be/tools/lora-calculator/

TABLE I: Expected battery lifetime (best and worst case) of LoRaWAN devices for different IoT applications operating on a coin cell (225 mAh).

Application	Average time between messages (s)	Payload size (bytes)	# Uplink Packets		Life Time	
			Best	Worst	Best	Worst
Roadway Signs	600	20	10 073	57 279	2 months 1 week	1 year 1 month
Traffic Signs	30	1	17 418	70 693	6 days 1 hour	3 weeks
Credit machine in grocery	120	24	9 330	63 790	1 week	2 months 3 weeks

TABLE II: Energy profile [5] used in the case.

State No.	State Description	Power (mW)	Duration (ms)
1	Sleep	5.7e-3	-
2	Processing	15	5
3	Tx prep.	12.5	40
4	Tx	Tab. III	Eq. 1
5a	Wait Rx 1	5.7e-3	1000
5b	Wait Rx 2	5.7e-3	1000 - len(state 7)
6	Rx prep.	8.25	3.4
7	Rx1	36.96	airtime(DR=DR_tx)
8	Rx2	34.65	airtime(DR=3)
9	Rx post proc.	8.3	10.7

2) Reducing the required transmit power at the node. The exploitation of (large) antenna arrays at the base station side is considered to improve reception sensitivity.

A. Escape Route 1: Cross-Layer Approach to minimize energy consumption

We have assessed the impact of the package length as a first cross-layer optimization opportunity.[2] As expected, the average energy consumption per payload byte decreases when sending larger packets (Fig. 3). To save energy, non-time-critical data can be accumulated, because by increasing the payload size

1) the overhead related to header information decreases,
2) the overhead of starting and initializing a transmission lowers,
3) the number of retransmissions in a stable propagation environment reduces,
4) the number of down-link receive windows is also lower.

[2]This escape route is introduced in a paper still under review.

TABLE III: Measured power [5] (drawn from battery) for the defined finite transmit power states.

Selected transmit power (dBm)	2	5	8	11	14
Power (mW)	91.8	95.9	101.6	120.8	146.5

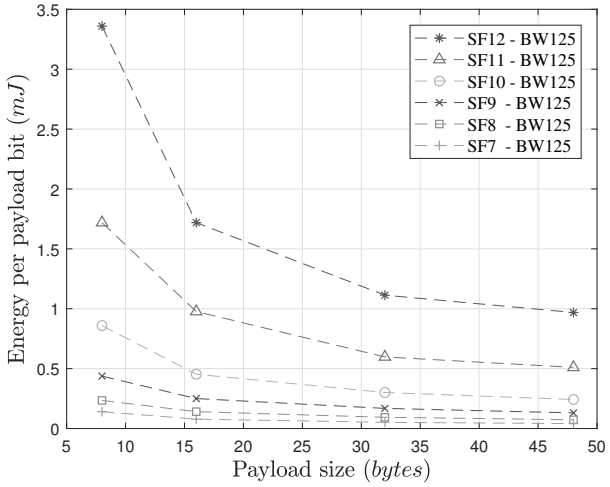


Fig. 3: Energy consumption per bit reduces when more data is being transmitted at higher data rates in LoRaWAN.

This observation has also been confirmed when simulating environments where collisions and retransmissions occur [7]. This conducted work is still under review.

Despite the aforementioned beneficial effects of increasing the payload size, sending more bytes per packet increases the total number of bytes which are sent sub-optimal. In LoRaWAN the Adaptive Data Rate (ADR) mechanism is defined to make a trade-off between range and energy consumption. Nodes closer to a gateway will be able to transmit at a higher rate, thereby reducing their airtime and thus energy. If nodes are located further away from the gateway, their data rate is reduced in order to improve the sensitivity of the receiver. Hence, it allows adapting the data rate to the channel characteristics. After receiving 20 uplink messages from the devices, the network will respond with the adequate parameters to accommodate for non-optimal propagation matched LoRa parameters.

For higher payload sizes this implies that more bytes have been sent before the LoRa parameters are adjusted to the channel. In addition, ADR changes the parameters in steps, yielding an even slower adaption to the propagation environment for larger payload sizes. This effect is clearly notable when observing the energy consumption over a short time period or when nodes have a slow data transmission rate. In quasi-static situations the impact will become negligible on the longer term. In dynamic situations, however, the trade-off on packet length may yield a different result.

To faster adapt to the channel, LoRa devices could first sent 20 smaller packets. This will result in reduced airtime and energy for packets which are sent with non-optimal parameters. A further in-depth investigation of packet length versus dynamics in the channel will be included in future work.

B. Escape Route 2: Exploitation of Antenna Arrays at the Base Station

A second possible escape route is increasing the number of antenna at the gateway. A promising technology, using a large number of antennas, is Massive MIMO (M-MIMO). In contrast to conventional Multi-User MIMO (MU-MIMO) systems, it is scalable to a large number of antennas and, thus, users or terminals. The main advantage of M-MIMO

over MU-MIMO is that the Base Station (BS) only has to learn the channel between the BS and the terminals.

As a consequence of the large number of service antennas, this technology is able to exploit spatial diversity through a favorable combination of propagation and processing aspects. Spatial multiplexing can be used to establish many parallel connections to the nodes in the same time-frequency slot. Notably, this operation expects that the channels, to the terminals, are adequately diverse. It was both theoretically and experimentally proven [8]–[10] that this hypothesis is reliable in various propagation environments, both in situations with direct Line of Sight (LOS) communication and in conditions with rich multi-path propagation. Furthermore, the array gain allows the nodes to reduce their transmit power due to an improvement of the sensitivity of the gateway.

The following example gives an indication of the expected energy gain. Consider a base station with 64 antennas. Theoretically, a reduction with a factor of $\sqrt{64}$ can be achieved by exploiting the array gain and channel hardening [8]. If 10% of the original Tx power is assumed still needed for pilot-based training, the total transmit energy can be reduced by a factor of four.

Despite the recent developments of Massive MIMO systems, a number of open challenges first need to addressed:

1) *Extend the coverage even when the BS is unable to hear the terminals.* M-MIMO operates in a two-phase closed loop. In the initialization phase, the channel response between the terminals and every individual antenna of the base station is determined via known orthogonal reference symbols, i.e. pilot symbols. In the communication phase, the terminals communicate through the same frequency-time slot. The coverage can only be extended when the terminals can be heard over large distances. However, when the base station is not yet trained (i.e. the channel is not known), and thus the benefits of Massive MIMO can not be applied, this coverage extension is not feasible.

   A possible solution is to do a "wide scan" through beamforming techniques [11]. This is not easily realized due to the inherently directivity of the antenna grid. An omni-directional grid can be approached by employing a circular grid. However, such a setup does not facilitate deployment in indoor or equivalent settings. In addition, a "wide scan" may be not feasible for terminals without a Line of Sight (LoS).
2) *Provide Medium Access Control for uncorrelated terminals with sporadic uplink-centric data.* The current mobile protocols are unable to support access of a large number of users due to the weight of signaling traffic originating from a large number of connections. Massive MIMO systems conceptually rely on measurements of the channel responses performed during the initialization phase by means of pilots. In Massive MTC, this needs to be realized for numerous (quasi) uncorrelated terminals. In a theoretical study, Fast Random Access [12] has been proposed. However, the challenge still exists on how this theoretical approach can be made practically feasible.
3) *Reduce the energy consumption at the terminals.* Despite the fact that an energy reduction has theoretically been demonstrated [8], it is far from obvious how to put the proposed approach into practice, taking the initial connection issues into account. In the past, it has been shown that a cross-layer approach is extremely suitable for optimizing the energy efficiency of the terminals [13]. In the case of Massive MIMO-based MTC, it will be probably necessary to employ a combination of PHY and MAC-layer approach to reduce the power consumption of the terminals.

IV. Conclusions and Future Work

As long range technologies such as LoRaWAN are being deployed to support a whole new range of IoT applications, it is clear that, when it comes to the autonomy of battery or environmentally powered devices, there is still room for improvement.

In this article we have presented a power-optimized design of an IoT node. This node's energy profile serves as a baseline that has been used in a simulation framework that allows for exploration and optimization of the energy consumption of LoRaWAN devices. We have explored 2 escape routes, that lead to increased autonomy:

1) cross-layer optimizations, such as accumulation of non-time critical data, minimize the energy consumption (per payload bit),
2) deployment of antenna arrays at the base station in conjunction with Massive MIMO techniques can further reduce the required transmit power, hence reducing the devices energy consumption.

Nevertheless, several challenges still remain. For example, a further in-depth investigation of packet length versus the dynamics in the channel will be included in future research. Additionally, coping with setup and medium access problems in a Massive MIMO approach poses a challenge as well. However, the exploitation of (large) antenna arrays at the base station side is definitively worthwhile, as it can drastically improve reception sensitivity, hence reducing the device's required transmit power.

References

[1] N. Sornin and A. Yegin, *LoRaWAN™ Specification*, LoRa Alliance Technical Committee Std., 2017, v1.1.

[2] R. Mallart, "IoT Solutions to a Telecom Paradigm Shift," 2016. [Online]. Available: https://static.tue.nl/fileadmin/content/faculteiten/ee/Onderzoek/Technologische_centra/Centre_for_Wireless_Technology/CWTe_2016_RR_SIGFOX-Mallart-IoTSolutions.pdf

[3] J. M. Meredith and I. Toufik, "Evolved Universal Terrestrial Radio Access (E-UTRA); NB-IOT; Technical Report for BS and UE radio transmission and reception," 3GPP, Tech. Rep. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3033

[4] Y. P. E. Wang, X. Lin, A. Adhikary, A. Grovlen, Y. Sui, Y. Blankenship, J. Bergman, and H. S. Razaghi, "A primer on 3gpp narrowband internet of things," *Comm. Mag.*, vol. 55, no. 3, pp. 117–123, Mar. 2017. [Online]. Available: https://doi.org/10.1109/MCOM.2017.1600510CM

[5] G. Ottoy, G. Leenders, and G. Callebaut, "LoRaWAN EFM32," doi: 10.5281/zenodo.1209414. [Online]. Available: https://github.com/DRAMCO/LoRaWAN_EFM32

[6] R. Huang, H. Li, B. Hamzeh, Y.-S. Choi, and S. Mohanty, *Proposal for Evaluation Methodology for 802.16p*, IEEE 802.16 Broadband Wireless Access Working Group Std., 2011.

[7] G. Callebaut, "LoRaWAN Network Simulator," doi: 10.5281/zenodo.1217124. [Online]. Available: https://github.com/GillesC/LoRaEnergySim/tree/v0.1.0

[8] T. Marzetta, E. Larsson, H. Yang, and H. Ngo, *Fundamentals of Massive MIMO*, ser. Fundamentals of Massive MIMO. Cambridge University Press, 2016.

[9] J. Vieira, S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. Wong, V. Öwall, O. Edfors, and F. Tufvesson, "A flexible 100-antenna testbed for massive MIMO," in *Globecom Workshops, 2014*. IEEE, 2014, pp. 287–293.

[10] A. Nordrum, "5G researchers set new world record for spectrum efficiency," *IEEE Spectr.*, 2016.

[11] C. Shepard, A. Javed, and L. Zhong, "Control channel design for many-antenna mu-mimo," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '15. New York, NY, USA: ACM, 2015, pp. 578–591. [Online]. Available: http://doi.acm.org/10.1145/2789168.2790120

[12] E. Björnson, E. De Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2220–2234, 2017.

[13] G. Miao, N. Himayat, Y. G. Li, and A. Swami, "Cross-layer optimization for energy-efficient wireless communications: a survey," *Wireless Communications and Mobile Computing*, vol. 9, no. 4, pp. 529–542, 2009.

Wireless Channel Modeling for Low-altitude UAV Networks in Urban Environments

Jianqiao Cheng[1] Ke Guan[2] François Quitin[1]

(1) Brussels School of Engineering, Université libre de Bruxelles (ULB)
(2) State Key Lab of Rail Traffic Control and Safety, Beijing Jiaotong University
jianqiao.cheng@ulb.ac.be ke.guan.bjtu@qq.com fquitin@ulb.ac.be

Abstract

The use of unmanned air vehicles (UAVs) as "flying base stations" is gaining significant traction in the wireless communication society. Such UAV-based wireless networks could be quickly deployed to increase the range of a network or to replace a damaged infrastructure in emergency situations. One important aspect of UAV-based communication networks is understanding the UAV propagation channel, especially the delay spreads and angular spreads that dictate the performances of OFDM and MIMO systems. In this paper we investigate the delay and angular spreads of the air-to-ground wireless channels, based on a large amount of ray-tracing measurements. We show that, contrary to expectations, the delay spread and angular spreads at the MT are larger in LOS cases than in NLOS cases, which can be explained by the particular geometry of the air-to-ground channels. We also show that only the EOA spread is significantly impacted by the distance between the UAV and the MT.

1 Introduction

The use of unmanned air vehicles (UAVs) for military and civilian applications has been steadily increasing in recent years. One potential application of UAVs is to use them as "flying base stations", allowing to extend the coverage and capacity of existing networks. Such UAV-based networks are interesting for extending the range of a network, serving remote areas, or temporarily replacing cellular infrastructure in disaster scenarios (e.g. flood, huricane, etc.). UAVs have high agility and mobility, and can be a cost-effective alternative for installing a fixed cellular infrastructure in specific situations.

One important aspect of UAV-based communication networks is understanding the UAV propagation channel. However, there is little literature that covers modeling or measurements of air-to-ground wireless channels. A few papers cover some basic parameters, such as path loss or fading characteristics, but no attempt has been made at modeling the spatio-temoral wireless channel, including parameters such as delay spread or angular spreads of the air-to-ground channels. Such parameters are essential for modern communication systems, as they ultimately determine the performances of a multicarrier or multi-antenna wireless system.

In this paper, we investigate the propagation characteristics of an air-to-ground wireless channel in urban environments. The carrier frequency is chosen as 2.6 GHz (corresponding to LTE carrier frequencies), and we analyze the channel for line-of-sight (LOS) and non-line-of-sight (NLOS) situations by using a ray-tracing software. We focus on parameters such as delay spread and angular spreads, which can provide valuable information about the required signal processing algorithms and antenna array configurations. We also investigate the link between the different parameters and the distance between the UAV and the mobile terminal (MT).

The rest of the paper is organized as follows. Section 2 briefly reviews current literature

on channel measurement and modeling for air-to-ground channels. In Section 3, we describe the ray tracing simulations for the air-to-ground channel characterization of urban environments. Section 4 presents the method used to extract angular spreads and delay spread of the urban air-to-ground channel snapshots. Finally, the analysis of the simulation results is presented in Section 5.
Conventions: In this paper, azimuth and elevation are defined as shown in Figure 1. Azimuth is defined as the angle of the projection of the vector in the xy-plane w.r.t. the x-axis, and elevation is the angle between the vector and the xy-plane. The UAV is chosen as the transmitter, and the angles w.r.t. the UAV are denoted as angles-of-departure, while the MT is chosen as the receiver, and the angles w.r.t. the MT are denoted as angles-of-arrival.

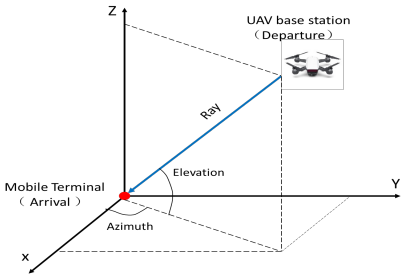


Figure 1: Representation of elevation and azimuth

2 Related Work

Some papers have investigated air-to-ground channel characteristic by field measurements and numerical simulations. In [1],[2] the authors use ray-tracing software to describe an urban air-to-ground propagation model. It is asserted that the probability of a UAV having an LOS link to the ground terminals is a function of both the UAV altitude and the elevation angle between the UAV and the terminals. The authors also propose a model to optimize the coverage radius of a UAV as a function of path loss and flying altitude. In [3], the author describe field measurements in a rural environment with a UAV flying at 5 different altitudes, and a path loss regression line is extracted. In [4], the air-to-ground channel model for a suburban city is investigated, and the correlations between the path loss, Ricean K-factors and root mean square delay spread are obtained.
While the angular characteristics of air-to-ground channels have not been investigated yet, angular channel measurements have been reported for situations where the transmitter is located high above street level (in or on top of a building). In [5], a measurement experiment was carried out to investigate the cluster characteristics of wideband 3D MIMO channels in outdoor-to-indoor scenario, and the channel model they proposed were identified jointly by the angles-of-arrival, angles-of departure and delays of individual multipath components. In [6], a measurement campaign was conducted using a 3D channel sounder where the transmitter was located at 120 positions from the 1st to the 4th floor of a building. The authors measured the distributions of elevation and azimuth angles for MT located in the streets. In [7], the same authors presented field measurements for an urban street canyon environment, with a focus on the fluctuation of the angular spreads when the MT moves along the street. The result in this paper showed that the distributions of angular spreads in 3-D channel are affected by both the distance between the transmitter and the MT, and the street layout surrounding the MT.

The objective of this paper is to propose a large set of simulated data (obtained using a ray-tracing software) to extract the spatio-temporal parameters of air-to-ground wireless channel. We will focus mostly on delay and angular spread, which characterize the distribution of power in the delay and angular domain, and eventually determine the performance of multicarrier and MIMO systems.

3 Simulation Setup

Our simulations are carried out by using CloudRT, the ray-tracing software developed by Beijing JiaoTong University [8]. The CloudRT software allows to simulate the direct ray between transmitter and receiver, first- and second-order reflections, diffracted rays along building edges and diffuse scattering on obstacles. We model a typical 3D urban city, shown in Figure 2, with 137 buildings with different heights going from 5 m to 70 m. For simplicity, we consider the surface of all buildings to be made of red brick. The total dimensions of the modeled terrain are 650m by 500m. The software simulates the wireless channel between 4 linear UAV trajectories (at a height of 120 m, shown in Figure 2) and 250 ground-based mobile terminals (MT) distributed uniformly over the whole map (at a height of 2 m, representing mobile phone users). The UAV is equipped with a downwards-facing patch antenna, while the ground receivers are equipped with vertically-oriented dipole antennas. The length of UAV trajectories is around 450m. The carrier frequency is set to 2.6 GHz (corresponding to the LTE carrier frequency) and the transmit power of the UAV transmitter is set to 0 dBm.

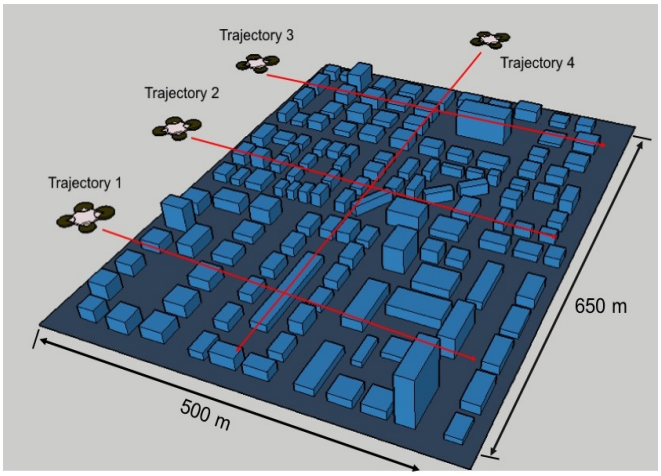


Figure 2: Ray-tracing simulation scenario

Our simulations are realized by fixing the 250 MT positions while varying the UAV positions along 4 predetermined trajectories (50 UAV positions per trajectory in total), as depicted in Figure 2. A total of $250 \times 50 \times 4$ snapshots were collected. For each snapshot, the different multipath components (MPC) between the UAV and MT were recorded. These were used to determine the delay spread, the Azimuth-of-Arrival (AOA) spread, the elevation-of-arrival (EOA) spread, the azimuth-of-departure (AOD) spread, and the elevation-of-departure (EOD) spread. As the direct path between UAV and MT can be obstructed by buildings, we have a collection of LOS and NLOS sitations. Out of the 50000 snapshots, we have 28626 LOS situations and 21374 NLOS situations.

4 Data Processing

The output of the ray-tracing simulation software provides us with all the propagation paths between a transmitter (the UAV) and a receiver (the MT), composed of the LOS path, first- and second-order reflections, diffracted rays and diffuse scattered paths. The delay spread σ_τ represents the dispersion of power in the delay domain, and is defined as [9]

$$\sigma_\tau^2 = \frac{\sum_{l=1}^{N} P_l \cdot (\tau_l - \tau_\mu)^2}{\sum_{l=1}^{N} P_l} \tag{1}$$

where P_l is the power for the l-th propagation path, τ_l is the delay of the l-th path, and τ_μ represents the mean delay that can be calculated as

$$\tau_\mu = \frac{\sum_{l=1}^{N} P_l \cdot \tau_l}{\sum_{l=1}^{N} P_l} \tag{2}$$

The delay spread is an important parameter, for example for multicarrier OFDM systems, as it determines the coherence bandwidth and the length of the cyclic prefix that is required for OFDM.
Similarly, the angular spreads represent the power dispersion in the angular domain, and can be defined separately for AOA, AOD, EOA and EOD. For example, the AOA spread σ_{AOA} is defined as [10]

$$\sigma_{AOA}^2 = \frac{\sum_{l=1}^{N} P_l \cdot (\angle\{e^{j(\phi_l - \phi_\mu)}\})^2}{\sum_{l=1}^{N} P_l} \tag{3}$$

where ϕ_l is the AOA of the l-th path, and ϕ_μ is the mean AOA defined as

$$\phi_\mu = \frac{\sum_{l=1}^{N} P_l \cdot \phi_l}{\sum_{l=1}^{N} P_l} \tag{4}$$

Angular spreads are a critical parameter for MIMO systems, as the ultimately determine which algorithm to choose (beamforming, spatial multiplexing, space-time block coding, etc.).

5 Result analysis

5.1 Delay Spread vs LOS Visibility

The cumulative distribution functions (CDFs) of the delay spread of the air-to-ground channel between the UAV and the MT in both LOS and NLOS scenarios are shown in Figure 3. It can be seen that in 90% of the cases, the delay spreads are below 21.5 ns and 23 ns for NLOS and LOS scenarios. In addition, the delay spread in LOS scenarios is slightly higher than in NLOS scenarios. This variation of measured delay spread between LOS and NLOS is quite interesting, as in typical scenarios the delay spread in NLOS is usually larger than in LOS. This counter-intuitive result will also be observed for angular spreads at the MT, and will be given below.

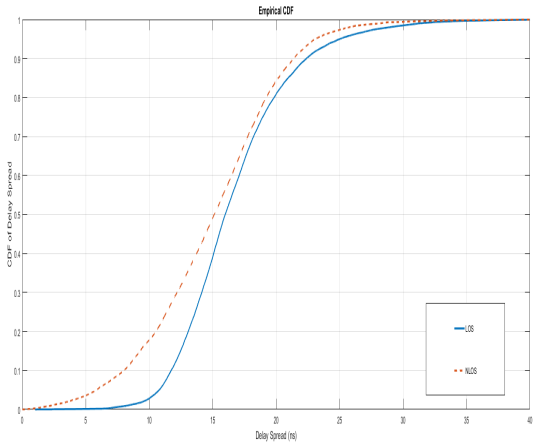


Figure 3: The CDF of Delay Spread

5.2 Angular Spread vs LOS Visibility

The angular spreads are important parameters to describe the spatial distribution of the incident power. Figure 4 plot the cumulative distribution of AOA/EOA spreads and AOD/EOD spreads. A first observation is that the AOA and EOA spreads in NLOS are higher than in LOS. This is again contrary to typical propagation scenarios where angular spreads are smaller for LOS than NLOS. The physical explanation for this will be explained below.

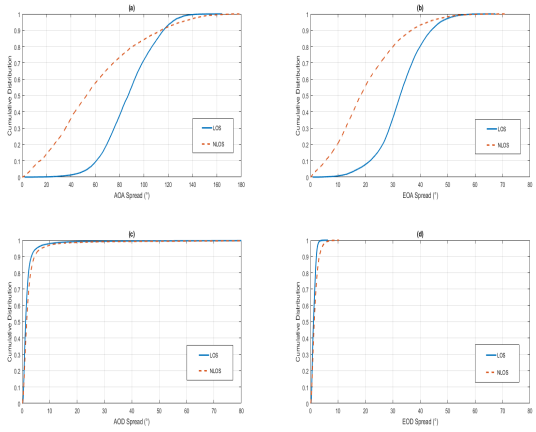


Figure 4: Angular Spread in LOS/NLOS

The AOD and EOD spreads concentrate in a very narrow range, no more than 15°, with no significant difference between LOS and NLOS. This indicates that the spatial dispersion of the power at the UAV side is very small, and that all the rays depart from the UAV in a very narrow range. This indicates that beamforming can be a simple strategy for focusing power towards the MT, or that highly directional antennas (or arrays) can be used to provide appropriate link performance between the UAV and MTs.

5.3 Why are delay spread and angular spreads at the MT higher in LOS than in NLOS?

In traditional propagation scenarios, the power of a LOS is so high that the mean delay (2) (or mean angle (4)) is approximately equal to the delay (or angle) of the LOS, thereby producing a small value for the delay (or angular) spread. From previous results, we can conclude that this is not the case for air-to-ground channels. To better understand why, we isolated a representative example from our simulation database. Figure 5 shows a top view of two snapshots for two successive UAV locations, one NLOS and one LOS. The LOS path in the second snapshot is shown in red, while the blue multipath indicate the reflected, diffracted and scattered multipath components. Our explanation below focuses on why the delay spread is higher for the LOS than for the NLOS case, but a similar argument can be made for AOA spread and EOA spread.

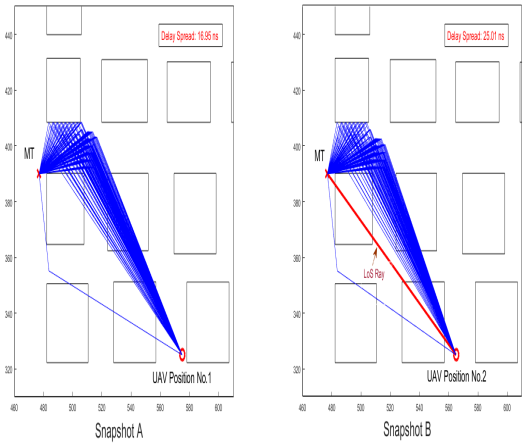


Figure 5: The multipath components for a NLOS and a LOS case. The red path represents the LOS path, while the blue paths represent all the other mulitpath components.

In the NLOS snapshot, it can be seen that most of the multipath are reflected from the building north of the MT, while there is no LOS path between the UAV and the MT. The mean delay is therefore roughly the propagation length from UAV to the building and from the building to the MT, and as all paths have roughly the same delay, the delay spread will be relatively small. In the LOS snapshot, there are two major contributions in the delay domain: 1) the LOS path, with a delay proportional to the UAV-MT distance, and 2) the reflected paths, with a delay proportional to the UAV-building-MT distance. The mean delay will be somewhere between these two major delays. The delay spread will then also be higher, since the power in the delay domain is spread between those two components.
This analysis, which can be generalized to most air-to-ground channel snapshots, can be summarized as follows. In NLOS situations, the particular geometry of the air-to-ground channels causes most of the power to come from a single cluster of scatterers located in the direction opposite the UAV, as shown in Figure 5. As soon as there is a LOS, the channel is composed of the same cluster of scatterers, as well as a LOS path coming from the direction of the UAV, causing the delay spread to increase. A similar argument can be made for the AOA spread and the EOA spread.

5.4 EOA spread vs UAV-MT Distance

We also investigated the delay and angular spread as a function of the distance between UAV and MT. Interestingly, we found that only the EOA spread shows a significant fluctuation with UAV-MT distance. In our simulations, the UAV-MT distance ranges from 118m to 675m. The cumulative distribution of the EOA spread under different distances for both LOS and NLOS scenarios are shown in Figure 6. It can be seen from Figure 6 that: (a) the EOA spread is decreasing when the UAV-MT distance is increasing in LOS scenarios, and (b) this trend is not distinct in NLOS scenarios.

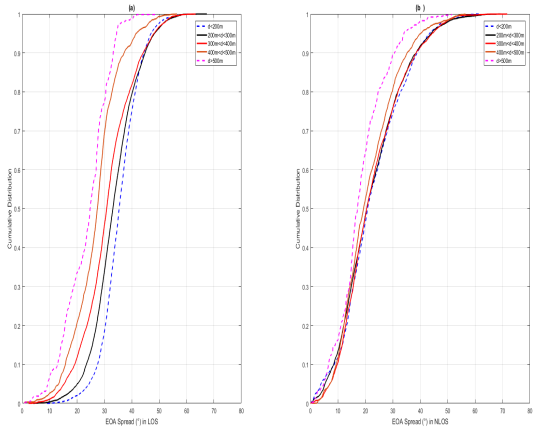


Figure 6: EOA Spread vs UAV-MT Distance

This correlation between EOA spreads and UAV-MT distances can be explained as follows. When the UAV and MTs are close (mostly LOS), the geometry is such that the paths from the UAV to the MT have a wide range of elevations (scattering from buildings, ground reflections, direct path). When the UAV is far from the MT (mostly NLOS), only a handful of paths that are reflected from buildings on one side of the UAV carry significant power, thereby strongly reducing the EOA spread.

6 Conclusion

In this paper, we present a detailed investigation of the wireless channel delay spread and angular spreads (at the UAV and the MT) as a function of LOS visibility and UAV-MT distance. A large set of ray-tracing simulations was performed, and 50000 snapshots were analyzed. It was observed that, contrary to expectations, the delay spread and angular spreads at the MT are larger for LOS than NLOS situations, which could be explained by the particular geometry of air-to-ground channels. Similarly, we observed that the EOA spread decreases when the distance between the UAV and the MT increases, which again can be explained thanks to the particular geometry of the air-to-ground channel.

References

[1] Al-Hourani, Akram, Sithamparanathan Kandeepan, and Abbas Jamalipour , "Modeling air-to-ground path loss for low altitude platforms in urban environments," Global Communications Conference (GLOBECOM), 2898 - 2904, 2014.

[2] Al-Hourani, Akram, Sithamparanathan Kandeepan, and Simon Lardner, "Optimal LAP altitude for maximum coverage,"IEEE Wireless Communications Letters 3.6, 569-572, 2014.

[3] Galkin, Boris, Jacek Kibida, and Luiz A. DaSilva, "Coverage analysis for low-altitude UAV networks in urban environments,"GLOBECOM 2017 - 2017 IEEE Global Communications Conference, 1 - 6,2017.

[4] Matolak, David W., and Ruoyu Sun, "Airground channel characterization for unmanned aircraft systemsPart I: Methods, measurements, and models for over-water settings," IEEE Transactions on Vehicular Technology 66.1,26-44, 2017.

[5] Detao Du; Jianhua Zhang; Chun Pan; Chi Zhang, "Cluster Characteristics of Wideband 3D MIMO Channels in Outdoor-to-Indoor Scenario at 3.5 GHz," Vehicular Technology Conference (VTC Spring), 1 - 6, 2014.

[6] Yalong Zhang; Ruonan Zhang; Stan X. Lu; Weiming Duan; Lin Cai, "Measurement and modeling of indoor channels in elevation domain for 3D MIMO applications." Communications Workshops (ICC), 2014 IEEE International Conference on. IEEE, 659 - 664,2014.

[7] Zhang, Ruonan, et al, "Measurement and modeling of angular spreads of three-dimensional urban street radio channels," IEEE Transactions on Vehicular Technology 66.5, 3555-3570, 2017.

[8] D. He, B. Ai, K. Guan, L. Wang, Z. Zhong, and T. Kuerner, "High-Performance Ray-Tracing Simulation Platform and Its Application to Mobile Communications," IEEE Communications Surveys and Tutorials, 2017.

[9] Ahmed M Al-SammanTharek Abd RahmanMarwan Hadri, "Path loss and RMS delay spread model for 5G channel at 19 GHz," Signal Processing its Applications (CSPA), 2017 IEEE 13th International Colloquium on,49 - 54, 2017.

[10] Ruonan Zhang; Xiaofeng Lu; Weiming Duan; Lin Cai; Jiao Wang, " Elevation domain channel measurement and modeling for FD-MIMO with different UE height," 2015 IEEE Wireless Communications and Networking Conference (WCNC), 70 - 75, 2015.

Distributed Edge-Variant Graph Filters

Mario Coutino, Elvin Isufi, Geert Leus*

Delft University of Technology, The Netherlands

E-mail: m.a.coutinominguez@tudelft.nl

Abstract

Graph filters are one of the core tools in graph signal processing. A central aspect of graph filters it to offer processing capabilities in a distributable manner. However, the filtering performance is traded with the distributed implementation if the communication and computational complexity needs to be limited. To tackle this issue, in this work, we generalize the state-of-the-art distributed graph filters to graph filter whose weights show a dependency on the nodes sharing information. We extend graph filters to filters where every node weights the signal of its neighbours with different values, while keeping the aggregation operation linear. This generalized graph filters yield significant benefits in terms of filter order reduction leading to amenable communication and complexity savings. In addition, we characterize the set of shift-invariant graph filters that can be described with edge-variant filters. For this purpose, we introduce a low-dimensional parametrization of such filters that allows a parsimonious definition of the elements of the family. This parametrization provides insights on the linear operator approximation through succession and composition of local operators, i.e., fixed support matrices, which span beyond the field of graph signal processing. The analytical and numerical results presented in this paper illustrate the potential and benefits of the general family of edge-variant graph filters.

Feasibility of Colonic Polyp classification with CNN based on Blue Light and Linked Color Imaging

R. Fonollà, F. van der Sommen, R.M. Schreuder, E.J. Schoon, P. H.N de With

Visual differentiation of benign and pre-malignant colonic polyps is an on-going challenge in clinical endoscopy routine. White Light Endoscopy (WLE) is the most common technique to visually assess lesions in the intestinal tract but is arguably unreliable due to hampering in polyp classification. Chromoendoscopy techniques are an alternative to enhance visual identification of gastrointestinal lesions by injecting a stain, improving differentiation between the mucosa, vessels and surface patterns in the intestinal tissue, but this requires additional chemical colorization to be injected in the body. In recent years, LED-based enhanced techniques like Blue Light Imaging (BLI) and Linked Color Imaging (LCI) are potentially promising alternatives to avoid the use of chemical stains and to obtain similar results. In addition, the need of in-vivo histopathology prediction would prevent costly resections and pathological examination for the benign polyps, which amounts to approximately 90% of all detected polyps [1].

In this work, a Convolutional Neural Network (CNN) is trained to automatically classify colorectal polyps between benign and pre-malignant tissue using three image acquisition modalities: White Light Imaging (WL), Blue Light Imaging (BLI) and Linked Color Imaging (LCI). The study involved a cohort of 60 patients from the Catharina Ziekenhuis, Eindhoven(CZE), The Netherlands. After histopathology examination, 49 patients were found with pre-malignant polyps and 11 with definitively benign polyps. Each polyp was acquired with three imaging modalities WL, BLI and LCI.

A first CNN was trained using the Kvasir Dataset[2] for automatic detection of gastrointestinal diseases. The dataset contains 8,000 images of 8 different classes ranging from healthy to malignant gastrointestinal tissue. The model was trained using the Adam optimizer, a learning rate of $5 \cdot 10^{-5}$ and a batch size of 64. The micro-averaged results showed a precision-recall of 95.66%. The second CNN was trained and fine-tuned using the CZE Dataset to classify each image as benign or pre-malignant. The results from the first CNN were used as an advantage to improve the training performance of the second network. A four-fold cross-validation was used to assess the performance of the model. The 180 images were divided in training (~75%) and validation (~25%) and 4 models were trained using each time a different validation dataset, thereby making sure that the same group of three images per patient was located either in the training or in the validation set. For each model, the same hyperparameters were used. The selected optimizer was RMSprop, the learning rate $2 \cdot 10^{-5}$, the batch size 20 and the decay rate $1 \cdot 10^{-3}$. A weighting coefficient of 2.9 was applied to all benign images to compensate for the imbalanced class. The mean average results of the four models are shown in Table 1.

Accuracy (%)	Sensitivity (%)	Specificity (%)	FPR (%)	FNR (%)
87.25±10.0	94.44±4.39	58.32±22.67	41.68±33.67	5.56±4.39

Table 1. Testing average metrics for the four CNNs models

The combination of WL, BLI, and LCI together with the achieved sensitivity results shows the feasibility of using CNNs to improve the classification of colonic polyps. In this work, we have built a robust model using a CNN to classify polyp malignancy from WL, BLI and LCI endoscopic images. This study is an improvement towards automatic assessment of colonic polyps and facilitates a future methodology to avoid expensive histopathologic assessment.

[1] W. B. Strum, "Colorectal Adenomas," *N. Engl. J. Med.*, vol. 374, no. 11, pp. 1065–1075, 2016.

[2] K. Pogorelov *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," *Proc. 8th ACM Multimed. Syst. Conf. MMSys 2017*, pp. 164–169, 2017.

On Constellation Shaping for Short Block Lengths

Y.C. Gültekin W.J. van Houtum F.M.J. Willems
Information and Communication Theory Lab
Signal Processing Systems Group
Department of Electrical Engineering, TU/e, The Netherlands
{y.c.g.gultekin, w.j.v.houtum, f.m.j.willems}@tue.nl

Abstract

Gaussian channel inputs are required to achieve the capacity of additive white Gaussian noise (AWGN) channels. Equivalently, the n-dimensional constellation boundary must be an n-sphere. In this work, constellation shaping is discussed for short block lengths. Two different approaches are considered: Sphere shaping and constant composition distribution matching (CCDM). It is shown that both achieve the maximum rate and generate Maxwell-Boltzmann (MB) distributed inputs. However sphere shaping achieves this maximum faster than CCDM and performs more efficiently in the short block length regime. This is shown by computing the finite-length rate losses. Then the analysis is justified by numerical simulations employing low-density parity-check (LDPC) codes of the IEEE 802.11 standard.

1 Introduction

We consider the transmission of information X over a one-dimensional AWGN channel under the average power constraint P. The capacity-achieving input distribution $P(x)$ for this channel is the zero-mean Gaussian with variance P. The loss in maximum achievable information rate (AIR) resulting from using a uniform $P(x)$ is called the shaping gap and is 0.255 bits per real dimension asymptotically in signal-to-noise ratio (SNR) and block length n. This loss can also be interpreted as the increase in required SNR to achieve a certain rate R and is 1.53 dB asymptotically in R and n. Fig. 1 illustrates this loss by plotting channel capacity and the maximum AIR for a uniform input distribution over $[-\sqrt{3P}, \sqrt{3P}]$.

On top of the shaping gap, there is also a performance degradation caused from the discrete nature of the practically used channel inputs X. Fig. 1 also illustrates this by plotting the mutual information (MI) between the channel input X and channel output Y for 2^m-ary amplitude shift keying (ASK) alphabets

$$\mathcal{X} = \{\pm 1, \pm 3, \cdots, \pm (2^m - 1)\}. \tag{1}$$

These curves stay close to the uniform capacity until $R = (m-1)$ and then converge to m asymptotically in SNR. In this paper, we consider ASK based transmission schemes while investigating several constellation shaping methodologies.

Constellation shaping can be defined as the optimization of channel input X with the purpose of decreasing the required average transmit power for a given target error probability. This topic is investigated from many perspectives ever since Shannon determined the capacity-achieving distribution for AWGN channels in his celebrated paper. For a detailed survey on signal shaping, see [1].

In this paper, two fundamentally different approaches for shaping will be examined. These are the recently-introduced constant composition distribution matching (CCDM) and the well-investigated sphere shaping. In Sec. 2, these methods are explained. In

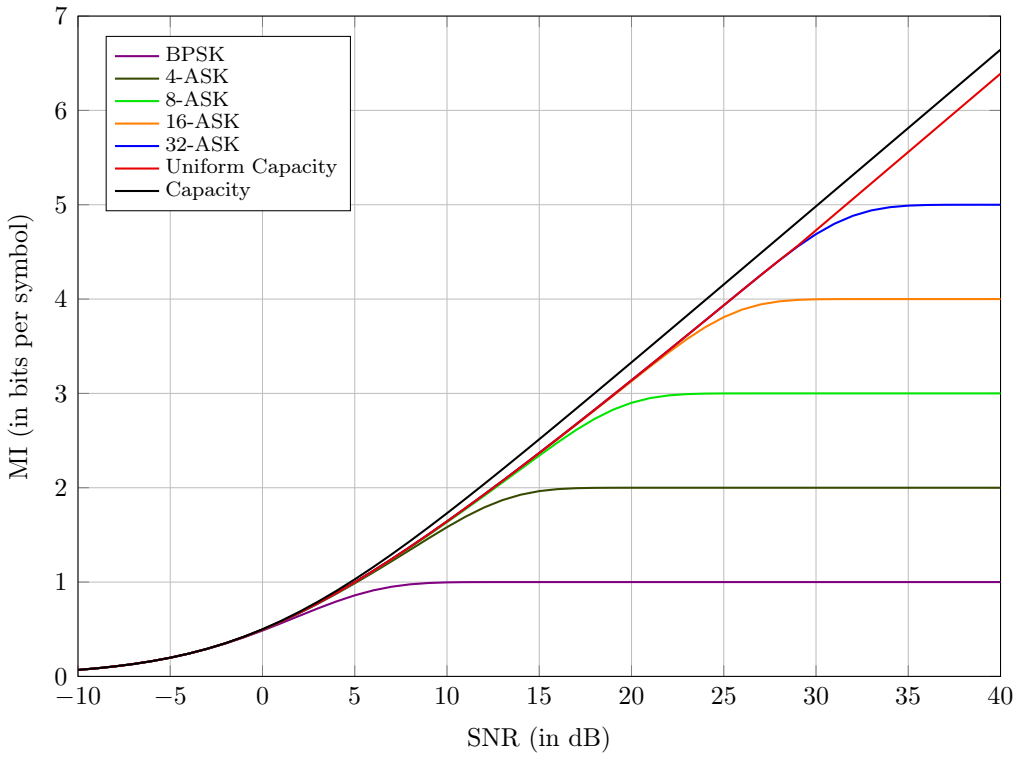


Figure 1: MI between channel input and output employing 2^m-ASK constellations for $m = 1, 2, \cdots, 5$ over AWGN channel. The Shannon capacity and the uniform capacity, i.e., the maximum AIR using uniform inputs, are also presented. The difference between the last two is 1.53 dB asymptotically, i.e., 0.255 bits per real dimension.

Sec. 3, an information-theoretical study is presented showing the asymptotic optimality of them. In Sec. 4, numerical results are presented to further illustrate the difference between the shaping approaches before the conclusion.

2 Shaping Approaches

Taxonomically, constellation shaping approaches are classified into two groups. The first one is probabilistic shaping (PS) where elements of $\mathcal{X}$ are employed with non-uniform probabilities. The second one is geometric shaping (GS) where positions of low-dimensional constellation points are optimized [1, Sec. 4.5]. We believe this categorization does not uncover the importance of sphere shaping where the boundary of the multidimensional constellation is structured in a way that will improve the performance. Although this again leads to non-uniform utilization of channel symbols as in PS, the indirect nature of accomplishing this motivates us to consider sphere shaping separately.

An information theoretically elegant way of constellation shaping is to first determine the capacity-achieving channel input distribution and then to obtain inputs with this distribution. Using channel input symbols with non-uniform probabilities according to a pre-defined distribution is called PS in this work. Recently, the concept of distribution matching (DM) is proposed to realize PS [2]. DM refers to any procedure that creates a non-uniform distribution that approximates, i.e., matches, a desired distribution. There are multiple ways of implementing distribution matchers

in the literature: Variable-to-fixed length prefix-free coding [2], arithmetic coding [3] etc. The approach in [3] is called CCDM and is combined with channel coding in the probabilistic amplitude shaping (PAS) framework in [4]. CCDM attracted a lot of attention especially in the optical communication society due to its high efficiency at large block lengths and the ability to create a fine granularity for the transmission rate, or equivalently, reach. We will use CCDM to represent PS here and discuss its advantages, disadvantages and efficiency for short to medium block lengths n.

From another perspective, the shaping gap can also be closed by sphere-shaping the multidimensional constellation boundary. As shown in [1] and [5], if an n-sphere is used as the boundary of the n-dimensional 2^m-ASK constellation* $\mathcal{X}^n$, the induced distribution on any low-dimensional constituent constellation approaches to a Gaussian asymptotically in n and m. Thus, the objective of transmitting Gaussian inputs can also be achieved by n-sphere shaping. Furthermore, achieving this goal by imposing a sphere constraint on an n-dimensional lattice might be more elementary from the channel coding perspective.

Next CCDM and n-sphere shaping will be explained.

2.1 Constant Composition Distribution Matching

Following the approach and notation in [3], we factorize $\mathcal{X}$ as $\mathcal{X} = \mathcal{A} \times \mathcal{S}$ where

$$\mathcal{A} = \{1, 3, \cdots, (2^m - 1)\}, \tag{2}$$

$$\mathcal{S} = \{-1, 1\}, \tag{3}$$

and limit our focus to n-amplitude sequences $A^n \in \mathcal{A}^n$.

Let P_A indicate a discrete distribution over the amplitude alphabet $\mathcal{A}$. The n-type distribution $P_{\bar{A}}$ which minimizes† $\mathbb{D}(P_A||P_{\bar{A}})$ is used to determine the composition $\#(a)$ as

$$\#(a) = nP_{\bar{A}}(a) \text{ for } a \in \mathcal{A}. \tag{4}$$

Then in a similar way to [7], arithmetic coding is utilized in [3] to implement a matcher that indexes all possible n-tuples $a_1 a_2 \cdots a_n$ having the same composition, i.e., constant composition. The functional diagram of such a matcher is given in Fig. 2. The rate of this matcher which indexes n-sequences with k bits is $R_{cc} = k/n$ bits per symbol.

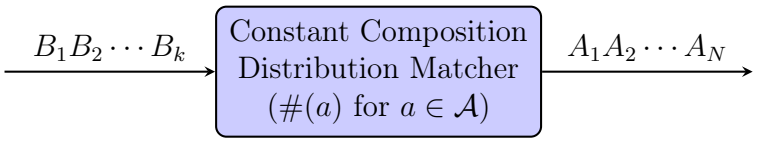


Figure 2: Constant composition distribution matcher. It maps k uniform bits (i.e., an index) to n amplitudes having a fixed composition $\#(a)$ for $a \in \mathcal{A}$. The mapping is invertible. Arithmetic coding is employed.

The distribution of these sequences at the output of the matcher is indicated by $P_{\tilde{A}^n}$. Geometrically, these sequences are located on the surface‡ of an n-sphere of radius

*Here the term 'n-dimensional ASK constellation' is used to indicate the n-fold Cartesian product of $\mathcal{X}$.

†For the optimum way of computing $P_{\bar{A}}$, see [6].

‡Note that there are multiple compositions that lead sequences to the same surface.

$r = \sqrt{E_{cc}}$ where

$$E_{cc} = \sum_{a \in \mathcal{A}} \#(a) a^2, \tag{5}$$

is the sequence energy. It is shown in [3] that the informational divergence between the desired and the output distributions

$$\mathbb{D}(P_{\bar{A}^n} || P_{\bar{A}}^n) = \mathbb{D}(P_{\bar{A}^n} || P_{\bar{A}}^n) + n\mathbb{D}(P_{\bar{A}} || P_A), \tag{6}$$

approaches zero for $n \to \infty$ which is equivalent to say for the rate of the constant composition code that

$$\lim_{n \to \infty} R_{cc} = \mathbb{H}(P_A). \tag{7}$$

Here it is assumed that the n-sequences are employed with equal probability. Note that letting $\mathcal{S}_{cc}$ indicate the set of these constant composition sequences, the rate can be written as

$$R_{cc} \triangleq \frac{\lfloor \log_2(|\mathcal{S}_{cc}|) \rfloor}{n} \text{ bits per symbol.} \tag{8}$$

Finally in [4], MB distributions are used as P_A to close the shaping gap relying on the fact that they have the maximum entropy for a fixed second moment, i.e., the variance[§] in this case [8].

The main advantages of CCDM are the asymptotic optimality and the virtual possibility of transmitting any rate by playing with P_A which can be invaluable in optical communications concerning the reach. On the other hand the disadvantages are the very long block length requirement to perform efficiently and inability to be parallelized. Note that the last is due to the use of arithmetic coding.

2.2 n-Sphere Shaping

Motivated by the asymptotic duality between n-sphere and Gaussian distribution, n-sphere shaping is the procedure of putting a maximum-energy constraint (i.e., the sphere constraint) on the possible n-sequences. The set $\mathcal{S}_\circ$ of these sequences can be defined as

$$\mathcal{S}_\circ = \left\{ a_1 a_2 \cdots a_n \,\middle|\, \sum_{i=1}^{n} a_i^2 \leq E_{\max} \right\}, \tag{9}$$

where $E_{\max}$ is the maximum sequence energy.

In the literature, the fact that a sphere constraint leads to a Gaussian distribution is proved using continuous approximation. Although there is no closed-form expression for the distribution that maximizes the AIR constrained by an ASK constellation, the reasoning in the continuous domain is extended to discrete domain somewhat pragmatically. Though in a coding setup, this makes sense since a shaping code can easily be combined with a (systematic) error-correcting code. In Sec. 3, we will show that the sphere constraint induces a MB distribution asymptotically when imposed on an n-dimensional ASK lattice.

[§]Here we implicitly assume that the distribution of signs will be uniform and $\mathcal{X}$ will have zero mean.

2.2.1 Enumerative Sphere Shaping

Although it is not necessary to specify an encoding strategy for n-sphere shaping to analyze its performance, here we outline enumerative sphere shaping (ESS) for the sake of completeness. Proposed in [9], ESS specifies efficient encoding and decoding algorithms to index energy-bounded sequences. These sequences are sorted lexicographically. In the context of this work, an enumerative shaper can be regarded as a black box, see Fig. 3. For a detailed discussion, see [5] and [9].

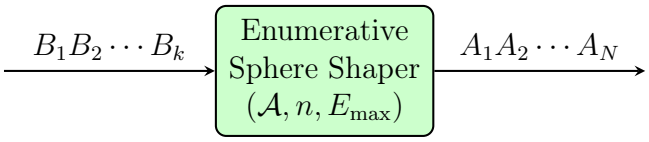


Figure 3: Enumerative sphere shaper. It is an invertible mapping from an index (i.e., k bits) to n amplitudes. All possible sequences have an energy no greater than $E_{\max}$.

Geometrically, energy bounded sequences are located in and on the surface of an n-sphere of radius $r = \sqrt{E_{\max}}$. The rate of this code is $R_\circ = k/n$ bits per symbol given that $E_{\max}$ is large enough to enclose more than $2^{nR_\circ}$ n-sequences. This rate can also be written as

$$R_\circ \triangleq \frac{\lfloor \log_2 (|\mathcal{S}_\circ|) \rfloor}{n} \text{ bits per symbol.} \tag{10}$$

Recently in [5], ESS is combined with non-systematic convolutional coding to improve the performance of IEEE 802.11 for $n = 96$.

Here we note that two different addressing algorithms for sphere shaping are given in [10]. The first algorithm is very similar to ESS where the only difference is sequences being sorted with respect to the n-dimensional shell that they are located on. The second algorithm is the well-known shell mapping which is introduced in [11] and included in the V.34 modem standard for $n = 16$ [12].

In the next section, we will investigate the asymptotic properties of constant composition and sphere codes.

3 Information-Theoretical Analysis

Let $\mathcal{C}$ be a code which consists of amplitude sequences $\boldsymbol{a}_k = a_{k,1}a_{k,2}\cdots a_{k,n}$ for $k = 1, 2, \cdots, L$. Here L indicates the number of codewords of length n in the code. All codewords occur with probability $1/L$.

The operational rate of such a code is $\log_2(L)/n$ bits per symbol and its operational average symbol energy is $\frac{1}{L}\sum_{k=1}^{L}\frac{1}{n}\sum_{i=1}^{n} a_{k,i}^2$.

Definition 3.1. (Achievability) The rate-energy pair (R, E) is called achievable if for each $\epsilon > 0$, for all n large enough, there exists a code with operational rate and operational average symbol energy satisfying

$$\frac{\log_2(L)}{n} \geq R - \epsilon, \tag{11}$$

$$\frac{1}{L}\sum_{k=1}^{L}\frac{1}{n}\sum_{i=1}^{n} a_{k,i}^2 \leq E + \epsilon. \tag{12}$$

Finally we define the rate-energy function as follows:

$$R(E) \stackrel{\Delta}{=} \max\{R : (R, E) \text{ is achievable}\}. \tag{13}$$

Theorem 1. The maximum achievable rate for average symbol energy E is

$$R(E) = \max_{A:\mathbb{E}[A^2]\leq E} \mathbb{H}(A).$$

The proof consists of a converse part and the corresponding achievability proof.

3.1 Converse

Consider a code. Assume that amplitude A_i for $i = 1, 2, \cdots, n$, has marginal distribution P_i generated[¶] by the uniform distribution over the codewords. Now the operational rate can be upper-bounded as

$$\begin{aligned} \frac{\log_2(L)}{n} &= \frac{1}{n}\mathbb{H}(A_1 A_2 \cdots A_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{H}(A_i|A_{i-1}, \cdots, A_1) \\ &\stackrel{(a)}{\leq} \frac{1}{n}\sum_{i=1}^{n} \mathbb{H}(A_i) \stackrel{(b)}{\leq} \mathbb{H}(\tilde{A}), \end{aligned} \tag{14}$$

where (a) follows from the fact that conditioning cannot increase entropy. If we say that $\tilde{A}$ is a random variable over alphabet $\mathcal{A}$ with distribution

$$\tilde{P}(a) = \frac{1}{n}\sum_{i=1}^{n} P_i(a), \tag{15}$$

then (b) is due to the convexity of entropy.

Next we observe that

$$\frac{1}{L}\sum_{k=1}^{L}\frac{1}{n}\sum_{i=1}^{n} a_{k,i}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} P_i(a)a^2 = \sum_{a\in\mathcal{A}} \tilde{P}(a)a^2 = \mathbb{E}[\tilde{A}^2].$$

We now conclude that for an achievable rate-energy pair (R, E), for all $\epsilon > 0$ and all large enough n, there exists a random variable A over $\mathcal{A}$ such that both

$$R \leq \frac{\log_2(L)}{n} + \epsilon \leq \mathbb{H}(A) + \epsilon, \tag{16}$$

$$E \geq \frac{1}{L}\sum_{k=1}^{L}\frac{1}{n}\sum_{i=1}^{n} a_{k,i}^2 - \epsilon = \mathbb{E}[A^2] - \epsilon. \tag{17}$$

If we let $\epsilon \downarrow 0$ we obtain that

$$R(E) \leq \max_{A:\mathbb{E}[A^2]\leq E} \mathbb{H}(A). \tag{18}$$

[¶]More precisely, for $a \in \mathcal{A}$, $P_i(a)$ is te number of codewords $\boldsymbol{a}_k$ for which $a_{k,i} = a$ divided by L.

3.2 Achievability Part

3.2.1 Achievability Based on Constant Composition Codes

Fix an energy E and assume that random variable A^* maximizes the entropy $\mathbb{H}(A^*)$ while satisfying the energy constraint $\mathbb{E}[(A^*)^2] \leq E$. Denote by $\{P^*(a), a \in \mathcal{A}\}$ the (MB) distribution corresponding to this random variable. For all large enough n, we now take a composition $\#(a)$ that satisfies

$$|\#(a) - nP^*(a)| \leq 1, \tag{19}$$

where $\#(a) \geq 0$ for all $a \in \mathcal{A}$, and $\sum_{a\in\mathcal{A}} \#(a) = n$.

It can be shown that the probabilities $P^*(a) > 0$, see [8, Example 11.2.3]. Therefore the normalized composition $\{\#(a)/n, a \in \mathcal{A}\}$ approaches entropy $\mathbb{H}(P^*)$ for increasing n.

Now for fixed n, consider a code consisting of all sequences having the composition $\{\#(a), a \in \mathcal{A}\}$. It can be shown that the operational rate of this constant composition code approaches the entropy $H(\#(A)/n)$ of the normalized composition for increasing n, see again [8, Example 11.2.3], where Stirling approximation is used.

We conclude that the operational rate of the constant composition code approaches the entropy of the normalized composition, which approaches entropy $\mathbb{H}(P^*)$, for n large. Therefore $R(E) = \max_{A:\mathbb{E}[A^2]\leq E} \mathbb{H}(A)$ is achievable for all E.

3.3 Optimality of Sphere Codes

Definition 3.2. (Sphere Code) A code is a sphere code if there exist no sequences, not in the code, with energy smaller than the energy of a codeword.

Theorem 2. For each code with operational rate R and energy E, there is a sphere code with operational rate R_o and operational symbol energy E_o such that

$$R_o = R \text{ and } E_o \leq E. \tag{20}$$

Proof. Just replace codewords by sequences outside the code with lower energy until the code is a sphere code. □

Theorem 2 along with the optimality of constant composition codes leads to the conclusion that sphere codes achieve the maximum rate as well.

Note that the code $\mathcal{S}_\circ$ defined before in Section 2.2 is a sphere code. The enumerative sphere code with rate as in (10) need not be a sphere code since sequences are sorted lexicographically and the ones with index 2^{nR} or more are not used. Observe that in the first algorithm in [10], sequences with largest index also have the largest energy and therefore the remaining sequences form a sphere code.

3.4 Maxwell-Boltzmann Distribution and Comparison

The distribution that achieves maximum entropy under an energy constraint is called Maxwell-Boltzmann distribution, see [8]. It is easy to show that the maximum entropy distribution is unique. This follows directly from the strict convexity of the entropy function. Since sphere codes result in maximum entropy under an energy constraint, the corresponding average marginal distribution (15) approaches the MB distribution.

To show the superiority of sphere codes over constant composition codes for short block lengths, we compute the finite-length rate loss

$$R_{\text{loss}} = H(A) - R, \tag{21}$$

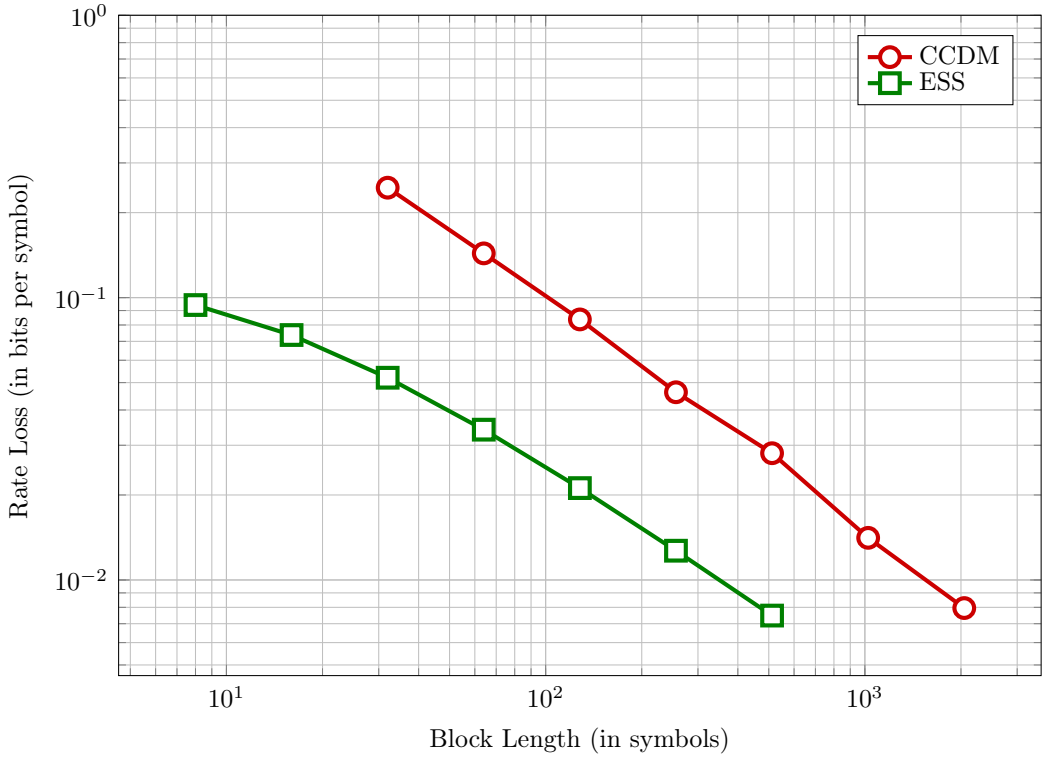


Figure 4: Block length (in symbols) vs. finite-length rate losses (in bits per symbol) of constant composition and sphere codes. The target rate is $R = 1.75$ for $\mathcal{A} = \{1, 3, 5, 7\}$.

where A is MB distributed and $\mathbb{E}[A^2]$ is equal to the average energy of the code used to achieve R. The results are presented in Fig. 4.

Here we fix the target rate $R = 1.75$, and find the composition $\#(a)$ and $E_{\max}$ that achieve R for constant composition and sphere codes respectively. It appears that for a target rate loss of 0.1 and 0.01 bits per symbol, constant composition codes require approximately 10 and 5 times larger block lengths than sphere codes, respectively. Furthermore, due to their definition, see Definition 3.2, sphere codes have the smallest block length requirement for a target rate loss.

3.5 Notes

We note that one counter argument may be the following: The shaping gain, i.e., the gain in average energy or in rate, does not necessarily imply a similar gain in AIR or an SNR improvement for a given error probability. Furthermore, utilization of constant composition codewords may enable better decoding performance than a sphere code in some cases. As an example, the constant composition property is exploited by a type check in successive cancellation list decoding of polar codes in [13].

A second note concerns the work presented in [14]. Here the shell mapping which is also an indexing method for n-sphere shaping is compared with CCDM. There are two fundamental differences between shell mapping and ESS: First, shell mapping sorts sequences with respect to their energies, i.e., the n-dimensional shell that they are located on, instead of lexicographical ordering. Second, the algorithm employs the divide and conquer principle (which requires multiplications) instead of operating in a sequential manner.

Finally, a bridge between sphere shaping and CCDM, namely partition-based distri-

bution matching (PBDM), is established in [15]. In this work, the constant composition constraint is relaxed by employing multiple compositions having the desired composition as the ensemble average. To put it simply, instead of considering a single shell, multiple nested shells are utilized. To select the corresponding composition, a variable-length prefix of the binary input is used which can be considered as a disadvantage. Although not as efficient as a sphere code, PBDM also achieves the maximum rate faster than CCDM.

4 Numerical Results

Monte Carlo simulations are used to compare the performance of constant composition and sphere codes in the short to medium block length regime. For CCDM simulations, PAS scheme is realized as in [4]. For sphere shaping simulations, the matcher in the PAS scheme is replaced by an enumerative sphere shaper. As the channel code, systematic LDPC codes of IEEE 802.11 [16] are employed with two different codeword lengths, i.e., $n_c = \{648, 1944\}$ bits. Note that the 2^m-ASK demapper on the receiver side assumes that the symbols are independent and identically distributed with either the n-type distribution of the constant composition code or the average marginal distribution of the sphere code. To achieve the target rate $R = 2.67$ based on 16-ASK, uniform transmission is simulated with the rate $r_c = 2/3$ code whereas shaped transmissions require $r_c = 3/4$.

Table 1: Parameters for Shaping

Approach	n	#(a) or $E_{\max}$	# seq.	E	Gain (in dB)
CCDM	162	(34, 32, 28, 23, 18, 13, 9, 5)	$2^{432.06}$	48.31	0.44
	486	(112, 103, 88, 69, 50, 33, 20, 11)	$2^{1296.15}$	42.22	1.02
ESS	162	796	$2^{432.13}$	39.74	1.29
	486	2326	$2^{1296.02}$	39.10	1.35

To obtain 2^{nR} n-sequences, the compositions and $E_{\max}$ values given in Table 1 are used for CCDM and ESS, respectively. Energy per symbol and shaping gain values of these schemes are also given in the same table. The shaping gain is computed with respect to the average energy equation $(2^{2H(X)} - 1)/3$ of uniform ASK constellations which is 53.42 for this target rate[‖].

In Fig. 5, the performances of the two different shaping techniques are presented. For $n = 162$, sphere shaping requires 0.85 dB less SNR than CCDM to achieve 10^{-3} frame error probability while it drops to 0.35 dB for $n = 486$. This justify our earlier statement that as n decreases, the difference between required block lengths increases in favor of sphere shaping.

5 Conclusion

Two different shaping techniques are discussed: Constant composition distribution matching and n-sphere shaping. It is shown that both asymptotically achieve the maximum rate and induce MB distribution. Sphere codes perform more efficiently (especially for short block lengths), i.e., approach the maximum rate (MB) faster than

[‖]Note that in this work $H(X) = R + 1$.

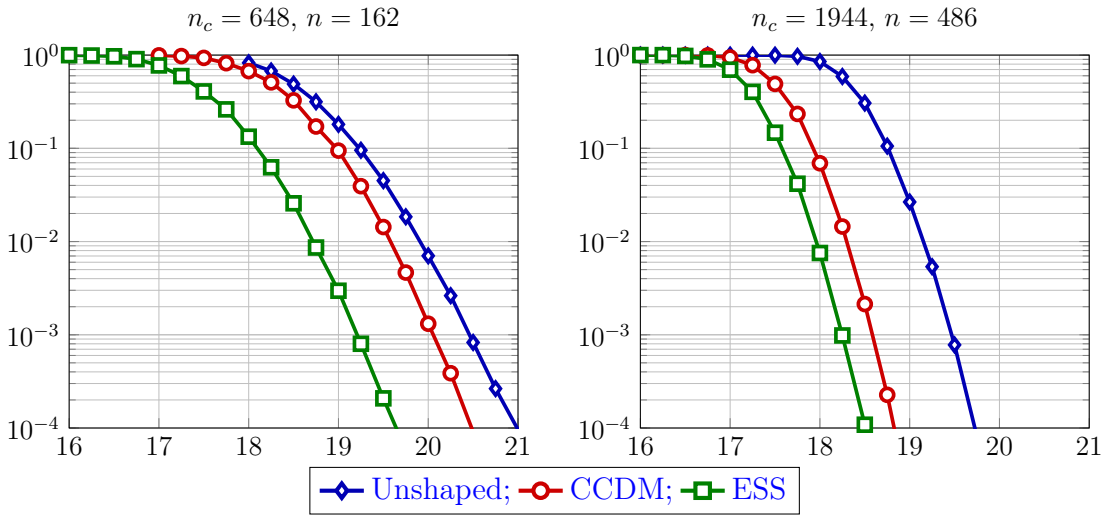


Figure 5: SNR vs. frame error probability for 16-ASK at target rate $R = 2.67$ for $n_c = \{648, 1944\}$ with and without shaping.

constant composition codes. This is shown by investigating the finite-length rate losses. Finally the comparison is justified by presenting Monte Carlo simulations employing LDPC codes.

References

[1] R. F. H. Fischer, *Precoding and Signal Shaping for Digital Transmission*. Wiley-Interscience, 2002.

[2] G. Böcherer and R. Mathar, "Matching Dyadic Distributions to Channels," in *2011 Data Compression Conference*, March 2011, pp. 23–32.

[3] P. Schulte and G. Böcherer, "Constant Composition Distribution Matching," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 430–434, Jan 2016.

[4] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth Efficient and Rate-Matched Low-Density Parity-Check Coded Modulation," *IEEE Trans. on Commun.*, vol. 63, no. 12, pp. 4651–4665, Dec 2015.

[5] Y. C. Gültekin, W. J. van Houtum, S. Şerbetli, and F. M. J. Willems, "Constellation Shaping for IEEE 802.11," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct 2017, pp. 1–7.

[6] G. Böcherer and B. C. Geiger, "Optimal Quantization for Distribution Synthesis," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6162–6172, Nov 2016.

[7] T. V. Ramabadran, "A Coding Scheme for m-out-of-n Codes," *IEEE Transactions on Communications*, vol. 38, no. 8, pp. 1156–1163, Aug 1990.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 1991.

[9] F. Willems and J. Wuijts, "A Pragmatic Approach to Shaped Coded Modulation," in *IEEE 1st Symp. on Commun. and Veh. Technol. in the Benelux*, 1993.

[10] R. Laroia, N. Farvardin, and S. A. Tretter, "On Optimal Shaping of Multidimensional Constellations," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1044–1056, Jul 1994.

[11] G. R. Lang and F. M. Longstaff, "A Leech Lattice Modem," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 6, pp. 968–973, Aug 1989.

[12] G. D. Forney, L. Brown, M. V. Eyuboglu, and J. L. Moran, "The V.34 High Speed Modem Standard," *IEEE Communications Magazine*, vol. 34, no. 12, pp. 28–33, Dec 1996.

[13] T. Prinz, P. Yuan, G. Böcherer, F. Steiner, O. İşcan, R. Böhnke, and W. Xu, "Polar Coded Probabilistic Amplitude Shaping for Short Packets," in *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2017, pp. 1–5.

[14] P. Schulte and F. Steiner, "Shell Mapping for Distribution Matching," *ArXiv e-prints*, Mar. 2018. [Online]. Available: http://arxiv.org/abs/1803.03614

[15] T. Fehenberger, D. S. Millar, T. Koike-Akino, K. Kojima, and K. Parsons, "Partition-Based Distribution Matching," *ArXiv e-prints*, Jan. 2018. [Online]. Available: https://arxiv.org/abs/1801.08445

[16] "IEEE Standard for Information technology–Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks–Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, pp. 1–3534, Dec 2016.

On the Effect of Polarization for Reliable Massive MIMO Communication

Sara Gunnarsson[1,2], Jose Flordelis[1], Liesbet Van der Perre[1,2] and Fredrik Tufvesson[1]
[1]Department of Electrical and Information Technology, Lund University, Sweden
{sara.gunnarsson, jose.flordelis, fredrik.tufvesson}@eit.lth.se
[2]Department of Electrical Engineering, KU Leuven, Belgium
liesbet.vanderperre@kuleuven.be

Abstract

Massive MIMO is an important technology when developing future wireless systems. In a massive MIMO system, base stations are deployed with a large number of antennas and thereby spatial diversity can be exploited in order to increase reliability. Then the small-scale fading decreases and the channel starts to behave deterministically. This is called channel hardening. Here, reliability in massive MIMO systems, when also exploiting polarization diversity, is investigated.

We analyze channel measurements of a 128-port cylindrical array and nine closely-spaced users in an indoor auditorium. The array consists of 64 dual-polarized patch antennas and the users have omni-directional antennas with vertical polarization. User antennas are tilted 45 degrees and follow some small movements.

The measurements show a considerable impact of polarization on reliability in a massive MIMO system. We demonstrate what to expect in terms of variations of channel gain over the base station array, as well as the standard deviation of channel gain for each base station antenna in the array*. We analyze the variations between the two polarizations and how they complement each other. We observe that for most users, one polarization outperforms the other. This shows the importance of having both polarizations present in the array in order to have reliable communication. The channel gain for 64 antennas over frequency and time is also analyzed in comparison to a single antenna when using one polarization or both polarizations. Channel hardening, in terms of decrease of standard deviation of the experienced channel gain, is evaluated when increasing the number of base station antennas. This is done for each one of the polarizations or a combination of both, the latter being the scenario where the channel hardened the most, as seen in Fig. 1. Based on these results, the conclusion is that polarization diversity is an important factor to consider when aiming for increased reliability in a massive MIMO system.

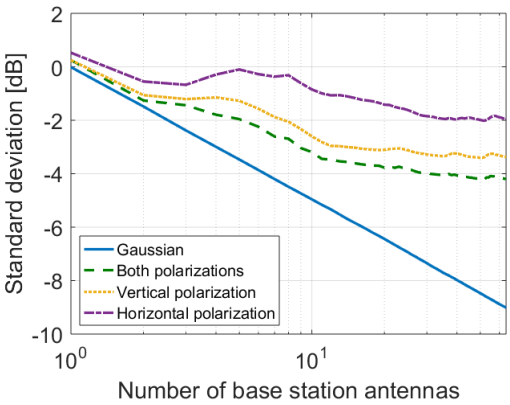


Figure 1: Standard deviation of channel gain versus number of base station antennas for one of the users.

*These results have also been accepted and will be presented at SPAWC 2018.

Distributed adaptive signal estimation in wireless sensor networks with noise in the exchanged signals

Fernando de la Hucha Arce[1], Marc Moonen[1], Marian Verhelst[2], Alexander Bertrand[1]
[1+2] KU Leuven, Dept. of Electrical Engineering (ESAT)
[1]STADIUS group, [2]MICAS group
Kasteelpark Arenberg 10, 3001 Leuven, Belgium
{fernando.delahuchaarce, marc.moonen, marian.verhelst, alexander.bertrand}
@esat.kuleuven.be

Abstract

Signal estimation in a wireless sensor network (WSN) aims to recover a desired signal from the noisy observations of a set of sensors deployed over a certain area. Distributed processing provides a division of the signal estimation task across the nodes in the network, such that said nodes need to exchange pre-processed data instead of their raw sensor observations. This is advantageous because it reduces the volume of data that needs to be exchanged, and wireless communications generally cost the node more energy than the local computations required to generate these pre-processed data [1].

The distributed adaptive node-specific signal estimation (DANSE) algorithm [2] relies on the exchange of optimally fused signals to achieve the performance of its centralized equivalent. However, additional noise can be present in these fused signals, such as noise introduced intentionally by quantization, targeting data reduction, or unintentionally by communication errors. In this case, the theoretical justification of the fusion rules in the traditional DANSE algorithm is no longer valid to guarantee the convergence of the algorithm.

We tackle the design of new fusion rules that take into account the power of this extra noise. These fusion rules are derived from an upper bound on the node-specific estimation cost functions, and result in a new "noisy"-DANSE algorithm, N-DANSE, which is guaranteed to converge to a unique point. The convergence point is optimal when the network has a star topology, which is a special case of a tree topology [3]. Additionally, these fusion rules produce almost no increase in computational complexity compared to the traditional DANSE algorithm.

Finally, numerical simulations show that N-DANSE slightly improves the performance of the DANSE algorithm, where the achievable performance improvement depends on the power of the local errors affecting the fused signals.

References

[1] G. Anastasi, M. Conti, M. Di Francesco, and A. Passarella, "Energy conservation in wireless sensor networks: A survey," *Ad Hoc Networks*, vol. 7, no. 3, pp. 537 – 568, 2009.

[2] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks – part I: Sequential node updating," *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5277 –5291, oct. 2010.

[3] ——, "Distributed adaptive estimation of node-specific signals in wireless sensor networks with a tree topology," *IEEE Trans. Signal Processing*, vol. 59, no. 5, pp. 2196–2210, May 2011.

Effect of Splitter & Combiner Non-Idealities in mm-wave Hybrid MU-MIMO System

Abhijeet Kanitkar[1], Steve Blandino[2,3] , Claude Desset[3], André Bourdoux[3], Sofie Pollin[2,3]
[1] Technische Universität Chemnitz, Chemnitz, Germany
[2] KUL ESAT-TELEMIC, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[3] Imec, Kapeldreef 75, B-3001 Leuven, Belgium
Contact author: steve.blandino@imec.be

Abstract

One of the approaches to meet the high throughput requirements of 5G is the utilization of millimeter wave (mm-wave) frequency bands which offer huge available bandwidth. The digital architecture proposed to achieve spatial multiplexing gain in the bands below 6 GHz bandwidth is impractical at mm-wave due to the increased hardware constraints, high signal processing complexity, different channel conditions and high cost. The hybrid analog-digital architectures for multi-user Multiple Input Multiple Output (MIMO) systems promise high capacity, high reliability and their implementation at mm-wave appears more realistic. However, the use of non-ideal components attenuates the signal, introduces distortions and degrades the final system performance. While hybrid architectures have been studied intensively, accurate RF models are still missing. Power splitters and power combiners are key components of a hybrid architecture and their characterization is often neglected. The paper focuses on the non-idealities introduced by splitters and combiners. We propose models extrapolated from the available measurement data and perform a joint digital/analog optimization by using different digital precoding strategies. We demonstrate that splitters and combiners need to be considered when designing a hybrid MIMO architecture. We estimate a degradation of 9 dB in EVM compared to an ideal architecture at SNR=25 dB. However, optimization of splitter and combiner operating points improve the system performance by 3 dB. Moreover we emphasize that system power constraints i.e. the use of non-linear power amplifiers (PAs), affect the operating point of the splitter.

1 Introduction

The fifth generation (5G) of wireless communication systems is expected to provide very high data rates to support the user demands for various applications. The underutilized frequency bands ranging from 30 to 300 GHz, millimeter wave (mm-wave) frequency bands, offer huge available bandwidth. As mm-wave frequencies have very small wavelength, it allows large number of antenna elements to be packed in a small area. This facilitates the realization of large antenna arrays at both Base Station (BS) and User Equipment (UE). Massive MIMO systems increase coverage, capacity and improve link quality by making the use of spatial multiplexing, beamforming and diversity gain. Therefore a natural evolution is the integration of Massive MIMO and mm-wave technologies to boost the overall system capacity.

The digital architecture proposed to achieve spatial multiplexing gain in the bands below 6 GHz bandwidth is impractical at mm-wave due to the increased hardware constraints, high signal processing complexity, different channel conditions and high cost [1].

Solutions at mm-wave have been implemented using fully analog architectures as proposed by 802.11 ad. However, the performance of the analog beamforming is considerably poor as compared to the digital beamforming for two main reasons; First, the analog beamforming is frequency flat and cannot be adaptive to the channel frequency variations which might be significant in an indoor applications. Secondly, fully analog solutions suffer from inter-stream interference if the users are not well separated [2], [3], [4].

The hybrid beamforming architectures share the beamforming operation between digital and analog domains. The hybrid beamforming is a promising solution since it lowers the system complexity while guaranteeing the system requirements. Different hybrid architectures realized by implementing a reduced number of RF chains minimize the hardware complexity by achieving a near optimal capacity [5].

One of the main differences between the digital architecture and hybrid architecture is the presence of splitters and combiners. The use of a phased antenna arrays require to replicate the input signal in several identical copies enabling the system complexity reduction. This can be implemented using splitters and combiners. However as we are going towards larger antenna arrays in mm-wave Massive MIMO implementation, the study of the impact of the non-idealities introduced by these components on the system performance becomes necessary. Though passive splitters and combiners are simple to implement, they introduce insertion losses in the system. The frequency selectivity degradation in SISO case can be compensated by using a receiver equalizer.

In multi-user systems, however, the transmitted signal is affected by inter-user interference which is minimized by precoding. In zero-forcing precoding, the beamforming matrix is designed to eliminate inter-user interference. The signal to be transmitted is multiplied by the inverse of the channel matrix and results in an equalized and interference-free channel to each user. However real splitters and combiners attenuate the signal at the edges of the desired bandwidth. The compensation of the distorted precoded signal reduces the power in the other sub-carriers and thus inversion of channel leads to an excessive power penalty at transmitter. The use of OFDM is one solution but its implementation is not practical due to peak to average power ratio (PAPR) and industries implement single carrier. Therefore an active splitter with larger bandwidth as compared to the desired system bandwidth, is required to minimize the signal attenuation at the edges. Moreover, a variable gain is needed to tune the transmitted power considering the system power constraints.

At the receiver, signals coming from an array of several phased antenna need to be combined into a single output. Like splitter, an active combiner is required to compensate the losses due to signal attenuation as well as to provide sufficient output power. The trade-off between gain and bandwidth of an active splitters and combiners limit the performance of these components in MU-MIMO system.

Although hybrid analog-digital beamforming seems to be a promising solution and a hybrid MIMO system has been validated with success over the air transmission [2], it is still not clear how practical measurements differ from ideal performance of the system, since variety of impairments in the RF transceiver, degrade the quality of the signal and a practical implementation exhibit significant performance degradation [7]. The hybrid analog-digital precoding system with emphasis on their modeling and radio-frequency (RF) losses have been considered in [8]. It has been observed that the performance of hybrid schemes significantly deteriorates when the realistic losses are taken into consideration. [9] studies the impact of residual transmit (Tx)-RF impairments on MIMO wireless communication system parameters such as antenna configuration and transmit power. The Tx-RF impairment distortions are strongly dependent on Tx-power used per RF chain. The residual Tx-RF impairments can be significantly reduced by choosing the appropriate Tx-power level. The mm-wave massive MU-MIMO system implemented using analog beamforming has been analyzed in [10]. The paper focus on the impact of non-ideal passive combiners on analog beamforming architecture. The non-ideal combiners have significant impact on analog

architecture implemented using large number of RF chains and degrades the system performance.

In our previous work [2], we have demonstrated the spatial multiplexing gain using digital processing using a hybrid MIMO testbed using 60 GHz RF phased array modules. We have observed a frequency selective distortion due to hardware non-idealities which was compensated by using digital precoding. However no insight was given in the non-ideality and in its impact on the performance when compared to an ideal hybrid architecture. In this paper we study the impact of the non-idealities introduced by an active power splitters and an active power combiners in a 60 GHz multi-user hybrid MIMO system. The reminder of the paper is organized as follows: Section 2 introduces the system transmission model. In section 3 we model the splitters and combiners. In section 4 the impact of the non-idealities on the system performance is studied and the optimization of splitters and combiners operating points is proposed. Section 5 gives the concluding remarks.

2 System Model

The hybrid MU-MIMO system considered in this paper consists of M_{RF} RF chains, equipped with M antennas at BS. Each RF chain is connected to M/M_{RF} antenna elements. This partially connected hybrid beamforming architecture can transmit $K \leq M_{RF}$ streams towards K different UEs. Each UE has N antennas connected to one single RF chain. We consider a single carrier system, similar to 802.11ad standard. The transmitter has digital precoding and analog beamforming capabilities to compensate the wideband channel frequency fluctuations caused by multipath propagation. An adaptive digital baseband precoding can be performed by estimating the channel for each UE. At the transmitter, time domain data symbols are transformed to frequency domain and a precoding is performed per sub-carrier. The precoded symbols are then transformed back to time domain and a cyclic prefix is added before the analog beamforming. The frequency domain representation of the transmitted symbol in each sub-carrier k can be written as:

$$\mathbf{x}[k] = \mathbf{F}_{\mathrm{A}}\mathbf{S}_{\mathrm{s}}[k]\mathbf{F}_{\mathrm{D}}[k]\mathbf{s}[k], \tag{1}$$

where $\mathbf{s}[k] \in \mathbb{C}^{K\times 1}$ are the data symbols at the subcarrier k such that $E[\mathbf{s}^H\mathbf{s}] = K$, while $\mathbf{x}[k] \in \mathbb{C}^{M\times 1}$ are the precoded symbols transmitted over the air. The symbols are precoded in digital domain using the frequency selective precoding matrix $\mathbf{F}_{\mathrm{D}} = [\mathbf{f}_{\mathrm{D}}^1, \mathbf{f}_{\mathrm{D}}^2, \ldots, \mathbf{f}_{\mathrm{D}}^K] \in \mathbb{C}^{M_{\mathrm{RF}}\times K}$. The analog precoding is implemented using an array of phase shifters which are represented by $\mathbf{F}_{\mathrm{A}} = [\mathbf{f}_{\mathrm{A}}^1, \mathbf{f}_{\mathrm{A}}^2, \ldots, \mathbf{f}_{A}^{M_{\mathrm{RF}}}] \in \mathbb{C}^{M\times M_{\mathrm{RF}}}$ which becomes block diagonal in a partially connected architecture. The matrix $\mathbf{S}_{\mathrm{s}} \in \mathbb{R}^{M_{\mathrm{RF}}\times M_{\mathrm{RF}}}$ represents the signal attenuation due the splitters. Considering the circulant property satisfied by using the cyclic prefix, the received signal $\mathbf{y}[k]$ at the UE after the application of the receiver combining vector $\mathbf{w}_{\mathrm{A}}^u \in \mathcal{C}^{N\times 1}$:

$$\mathbf{y}[k] = \alpha\mathbf{S}_{\mathrm{c}}[k]\mathbf{w}_{\mathrm{A}}^{u\,H}\mathbf{H}^u[k]\mathbf{x}[k] + \mathbf{S}_{\mathrm{c}}[k]\mathbf{w}_{\mathrm{A}}^{u\,H}\mathbf{n}[k], \tag{2}$$

where $\alpha = \frac{1}{||\mathbf{F}_{\mathrm{A}}\mathbf{F}_{\mathrm{D}}||}$ sets the transmitted power constraint, $\mathbf{H}^u \in \mathbb{C}^{N\times M}$ is the downlink mm-wave channel matrix towards the user u and $\mathbf{n} \in \mathbb{C}^{N\times 1}$ is the additive gaussian noise vector with co-variance $\sigma_n I$. The matrix $\mathbf{S}_{\mathrm{c}} \in \mathbb{R}^{1\times 1}$ represents the signal attenuation due the combiners necessary in a hybrid MIMO architecture. The signal attenuation is frequency selective since active components are considered. Then, in the digital domain, we observe an equivalent channel:

$$\mathbf{H}_{\mathrm{eq}}^u[k] = \mathbf{S}_{\mathrm{c}}[k]\mathbf{w}_{\mathrm{A}}^{u\,H}\mathbf{H}^u[k]\mathbf{F}_{\mathrm{A}}\mathbf{S}_{\mathrm{s}}[k], \tag{3}$$

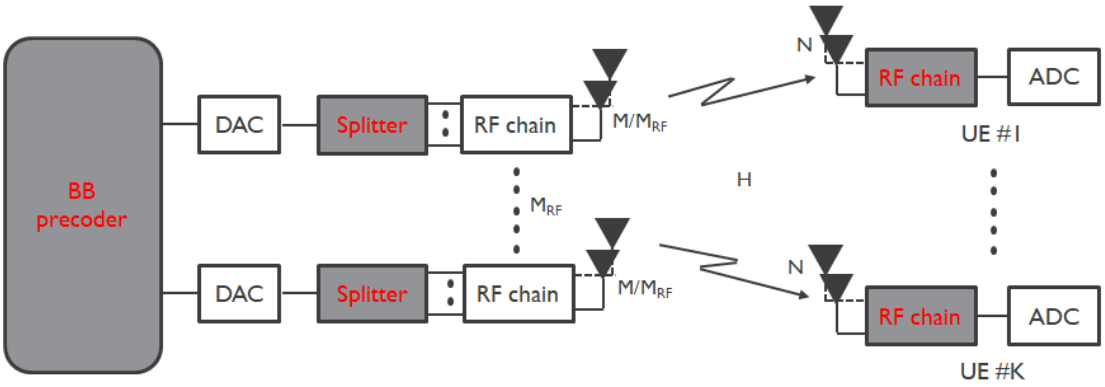


Figure 1: Multi user MIMO system with partially connected hybrid analog-digital beamforming architecture.

such that $\mathbf{H}_{\text{eq}}^{u} = [h_{u1}, h_{u2}, \ldots, h_{uM_{\text{RF}}}]$. The equivalent channel includes the effect of attenuation due to the splitters and combiners. The design of $\mathbf{F}_{\text{A}}$ is out of this scope of the paper.

The design of the digital precoding matrix $\mathbf{F}_{\text{D}}$ is based on the knowledge of the K channels $\mathbf{H}_{\text{eq}}^{u}[k]$ at the BS. We assume full knowledge of the equivalent channels at the transmitter but we consider a non-ideal channel estimation to take into account the effect of the splitters and combiners during the channel estimation procedure. Interference mitigation can be implemented by linear frequency selective precoding techniques like Zero Forcing (ZF) and Minimum Mean Squared Estimation (MMSE). The MMSE precoding matrix is given as:

$$\mathbf{F}_{\text{D}} = \mathbf{H}_{\text{eq}}^{H}(\mathbf{H}_{\text{eq}} \cdot \mathbf{H}_{\text{eq}}^{H} + \beta I)^{-1}, \tag{4}$$

where

$$\beta = \frac{K}{\rho}$$

K represents number of user and ρ is defined as a regularization factor. When $\rho \to 0$, $\beta \to \infty$, this results in Maximum ratio transmission (MRT) precoding which is usually not used in multi-user scenario as it does not provide inter-user interference. When $\rho \to \infty$, $\beta \to 0$, this results in ZF precoding.

3 RF modeling of Splitters & Combiners

This section focuses on the analysis of a hybrid MU-MIMO system by realistic modeling of the splitters and combiners. The testbed used for the modeling is as depicted in [2]. The mm-wave transceiver, denoted as RF chain in figure 1, is described in [11].

The BS has $M = 32$ antenna elements using $M_{\text{RF}} = 2$ transmitter units each of which having $M_{\text{sub}} = 16$ antenna elements. The BS serves to $K = 2$ UEs, each having a RF chain connected to $N = 4$ antenna elements.

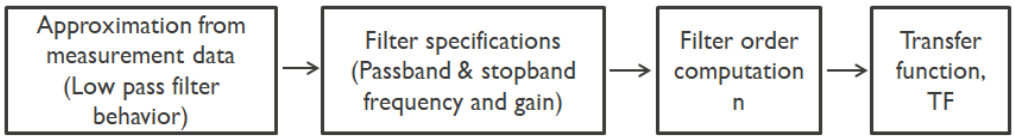


Figure 2: The algorithm used to generate splitter model from the available measurement data

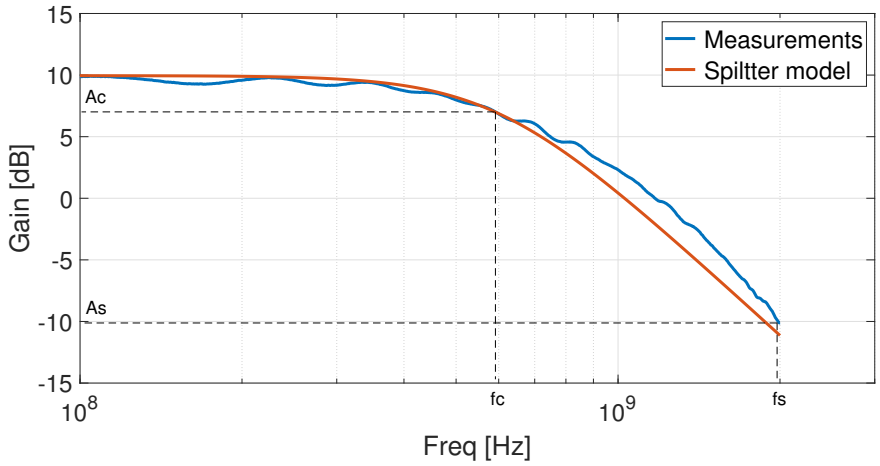


Figure 3: Second order butterworth LPF model accurately matches with splitter measured gain frequency distortion.

3.1 Splitter modeling

The transmitter has an external four ways splitter chip and an integrated four ways splitter. The external splitter is realized in 28 nm CMOS. The signals are split hierarchically in a tree of two levels of splitters. The splitter has two stages of gain controls g_1, g_2 implemented using programmable degeneration resistors with eight possible levels. The splitter bandwidth is controlled by a common mode voltage V_{cm} with 32 possible levels which determine the amount of current at each stage of the splitter. The choice of the triplet $(g_1, g_2, V_{\text{cm}})$ allows to program the gain G, at each antenna path of a sub-array, between -3 dB and 12 dB and a variable bandwidth between 400 MHz and 1.5 GHz.

The S-parameters of a single antenna path are available from measurements for all internal gain setting, g_1 and g_2, at two different V_{cm} values, for a total of 128 measurement points. The gain-bandwidth product at these points is constant, $A \cdot f_c = 4.5{\times}10^9$ dB·Hz where A denotes the DC-gain and f_c is cut-off frequency. To describe the splitter gain frequency distortion of a sub-array for all the 2048 possible splitter configuration (g_1,g_2,V_{cm}), a splitter model is extrapolated assuming a constant gain-bandwidth product.

Figure 2 summarizes the algorithm used to extrapolate the splitter model from available measurement data. From the measurements we observe that the splitter has a Butterworth low pass filter (LPF) behavior. Based on this assumption, we extract the filter specifications. The passband gain A_c is defined as the gain corresponding to the -3 dB cut-off frequency f_c. In the stopband, the magnitude of the response approaches zero with a gain A_s at the minimum frequency of the stopband f_s. Table 1 summarizes the filter specifications for the measurements in Figure 3.

The generic squared magnitude function of a n^{th} order Butterworth LPF is:

Table 1: Butterworth Low Pass Filter Specifications

Cut-off frequency f_c	593.64 MHz
Stop-band frequency f_s	1989.4 MHz
Cut-off gain A_c	7 dB
Stop-band gain A_s	−10 dB

$$G(f) = \frac{A}{1 + \left(\frac{f}{f_c}\right)^{2n}}, \tag{5}$$

The filter order is computed from equation (5) assuming $G(f_s) = A_s$:

$$n = \frac{1}{2} \frac{\log_{10}\left(\frac{A - A_s}{A_s}\right)}{\log_{10}\left(\frac{f_s}{f_c}\right)} \approx 2, \tag{6}$$

Hence, the splitter can be described with a second order Butterworth LPF transfer function as:

$$TF = \frac{\omega_c^2}{(s + \omega_c)^2} A. \tag{7}$$

where, A is the DC-gain, f_c is the cut-off frequency, $\omega_c = 2\pi f_c$, $s = j2\pi f$. Assuming a constant gain-bandwidth product, we perform linear interpolation and extrapolation on designed model to obtain the gain and cutoff frequency for all splitter programmable parameters.

3.2 Combiner model

The MU-MIMO system described in this work has 4 antenna path receiver, implemented using a mm-wave transceiver chip, with beamforming and signal combining capabilities at baseband. The 2 stage internal combiner is followed by 2 stages of a variable gain amplification. The combiner has gain programmability from 0 to 32 dB. The overall conversion gain of the receiver can be varied from 28 to 62 dB by tuning the gain setting of a trans-impedance amplifier (TIA) located after the local oscillator and a mixer.

Figure 4 depicts the conversion gain of the receiver with combiner gain is set to 30 dB and TIA gain setting is varied. The receiver model is generated using the same algorithm described for the splitter modeling. The butterworth 4^{th} order LPF model closely matches with the measurement data whose transfer function is given as:

$$TF = \frac{\omega_c^4}{(s + \omega_c)^4} A. \tag{8}$$

where, A is the DC-gain, f_c is the cut-off frequency, $\omega_c = 2\pi f_c$, $s = j2\pi f$.

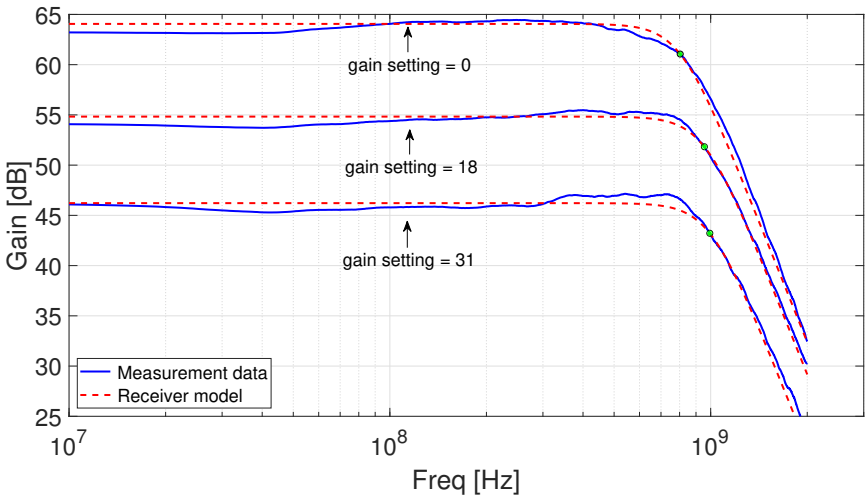


Figure 4: Fourth order butterworth LPF model accurately matches with receiver measured gain frequency distortion.

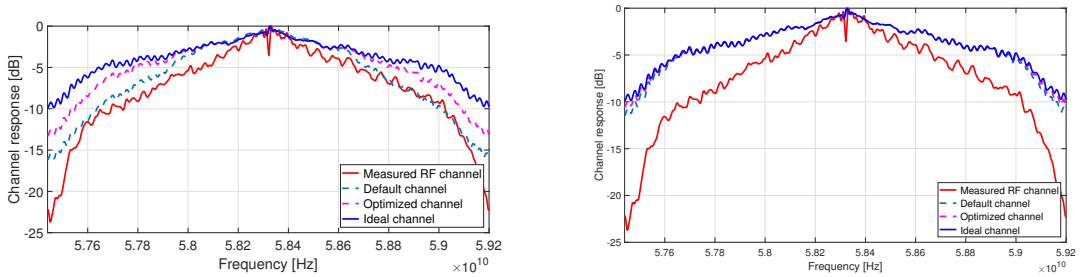
4 Impact of splitter/combiner non-idealities on MU-MIMO system

The impact of the splitter and combiner non-idealities are studied by integrating the designed models into a MIMO Matlab simulator. We transmit uncoded 16-QAM modulated symbols. The digital channel used for simulations is generated from a quasi-deterministic channel model in an indoor environment with 10° angular separation between UEs and BS & UEs are 2.4m apart [12].

Figure 5 compares the channel response of an ideal splitters and combiners with the measured RF channel which contains the non-idealities from all analog components. The degradation in channel is less for the optimized splitter and combiner as compared to the default operating point. The comparison shows that splitter with lower order LPF model results in more non-idealities and significantly deteriorates the signal as compared to the receiver impairments. This is due to the fact that in our testbed an external splitter is required to split the signal into 16 antenna paths whereas combiner consists of 4 antenna path and only an integrated combiner is used. Moreover the receiver combiner has higher cut-off frequency and hence has less impact on the signal attenuation. This leads to an important conclusion that the splitter and combiner non-idealities may becomes more critical as we are going towards larger antenna arrays and even more than 16 antenna paths might be needed in Massive MIMO implementation.

Moreover in a system with small separation between BS and UE, as in indoor applications, the receiver has a sufficient power. Therefore at the receiver, the focus is given only on the bandwidth optimization. The optimized combiner configuration is selected for a TIA gain setting which provides the highest bandwidth. On the other hand, splitter optimization is carried out by considering both gain and bandwidth. First the required splitter gain is determined considering the system power constrains and splitter internal gain parameters, g_1,g_2, are optimized. There are several combinations of (g_1,g_2) which resulted in the desired splitter gain. From these combinations, splitter bandwidth is then optimized by selecting V_{cm} corresponding to the highest cutoff frequency.

Figure 6 shows the effect of component optimization for different precoding techniques. The optimized components along with MMSE precoding improve the EVM by more than 3dB for high SNR. However, real models saturate the EVM at even higher



(a) Channel estimation under splitter model (b) Channel estimation under receiver model

Figure 5: Impact of splitter and combiner non-idealities on system performance.

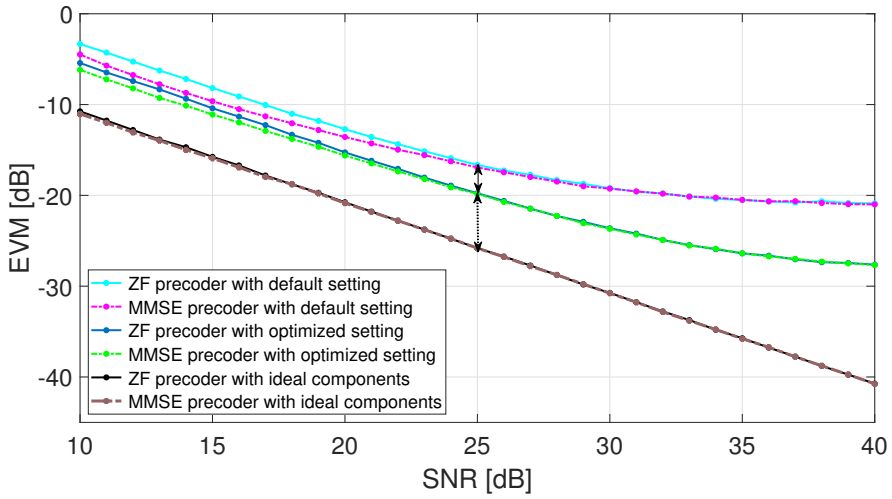


Figure 6: Hybrid MU-MIMO system with M=32:BS antennas, M_{RF}=2:BS RF-chains , K=2:UE, N=4:antennas per UE. At SNR=25dB, 9dB degradation in EVM compared to an ideal architecture. Optimization of splitter and combiner operating points improve the performance by 3dB.

SNR and introduce more than 6 dB decreased EVM compared to an ideal behavior. Although difference between MMSE and ZF is not significant at high SNR region, MMSE outperform ZF for optimized setting.

However in practice, thorough analysis of the entire integrated device is necessary. The mm-wave transceiver chip considered in this work contains the nonlinear PAs. The performance of the system depends on the signal to noise and distortion ratio SNDR. The presence of the non-ideal splitters reduce the signal power by a constant factor which is determined by the splitter operating point. The splitter gain determines the transmitter output power, hence it sets the received signal strength at each UE. Lower splitter gain degrades the transmitter signal while larger splitter gain is subject to smaller available bandwidth and thus add more distortions.

Assuming an analog RF input power of P_{in}, the transmitter output power is given as:

$$P_{\text{out}} = P_{\text{in}} G G_{\text{c}}. \tag{9}$$

where the gain G is set by the triplet $(g_1, g_2, V_{\text{cm}})$. The measured conversion gain for one antenna path, from input baseband to output RF, is $G_{\text{c}}|_{\text{dB}} = 30\,\text{dB}$ [11].

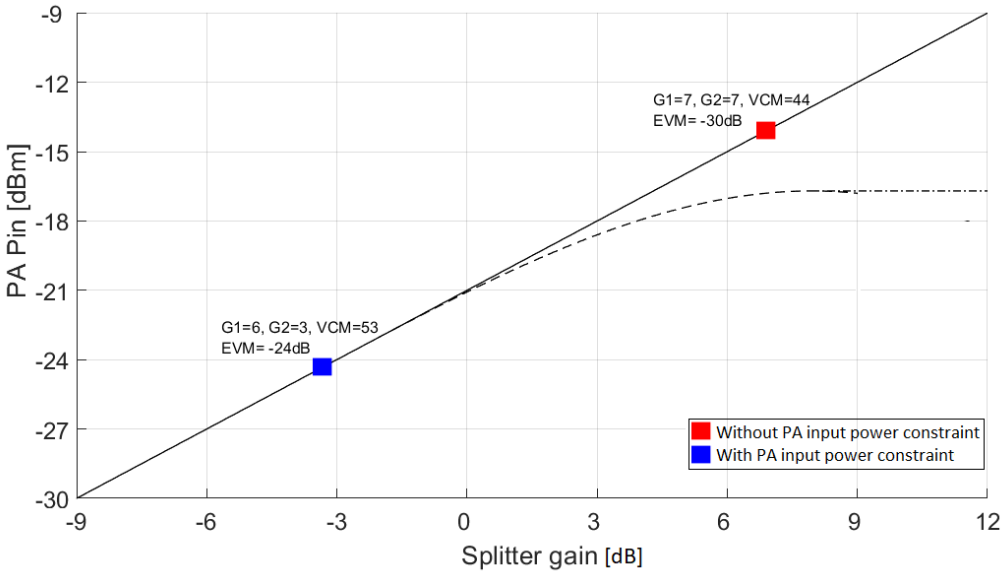


Figure 7: PA input power as a function of splitter operating point.

Therefore the splitter operating point $(g_1, g_2, V_{\text{cm}})$ need to be optimized considering the PA saturation level. The dependency of the per-antenna PA input power $P_{\text{in}}^{\text{PA}}$ as a function of the splitter gain is depicted in Figure 7 where $P_{\text{in}} = -21\,\text{dBm}$. Without any PA input power constraint, the splitter can be operated with a maximum internal gain setting which provides EVM of -30 dB. On the other hand optimization with PA input power constraint reduces the EVM by 6 dB. In spite of the decreased EVM, the system operates in a linear region of PA and a better constellation is achieved without compression of the transmitted symbols.

5 Conclusion

In this paper we present the effect of non-ideal splitters and combiners in a hybrid multi-user MIMO system at millimeter wave frequency. We propose an active power splitter and power combiner models based on real measurements. Moreover we optimize the splitter operating point assuming that the transmitter is using a non-linear PAs.

The simulations reveal that a hybrid MIMO system under realistic models exhibits a significant performance degradation compared to an ideal architecture. We estimate a degradation of 9 dB in EVM compared to an ideal architecture at SNR=25 dB. However, optimization of splitter and combiner operating points improve the system performance by 3 dB.

Understanding the impact of non-idealities emphasizes the need of RF modeling of components when designing the mm-wave hybrid MIMO architecture to make the system more power efficient. Larger antenna systems might require the use of more stages of active components introducing more distortion hence, the scaling of the performance with the number of antennas need to be addressed.

References

[1] Bogale, Tadilo Endeshaw, and Long Bao Le. "Beamforming for multiuser massive MIMO systems: Digital versus hybrid analog-digital." In Global Communications

Conference (GLOBECOM), 2014 IEEE, pp. 4066-4071. IEEE, 2014.

[2] Blandino, Steve, Claude Desset, Cheng-Ming Chen, Andre Bourdoux, and Sofie Pollin. "Multi-User Frequency-Selective Hybrid MIMO Demonstrated Using 60 GHz RF Modules." arXiv preprint arXiv:1711.02968 (2017).

[3] Alkhateeb, Ahmed, Jianhua Mo, Nuria Gonzalez-Prelcic, and Robert W. Heath. "MIMO precoding and combining solutions for millimeter-wave systems." IEEE Communications Magazine 52, no. 12 (2014): 122-131.

[4] Gimenez, Sonia, Sandra Roger, Paolo Baracca, David Martn-Sacristn, Jose F. Monserrat, Volker Braun, and Hardy Halbauer. "Performance evaluation of analog beamforming with hardware impairments for mmW massive MIMO communication in an urban scenario." Sensors 16, no. 10 (2016): 1555.

[5] Molisch, Andreas F., Vishnu V. Ratnam, Shengqian Han, Zheda Li, Sinh Le Hong Nguyen, Linsheng Li, and Katsuyuki Haneda. "Hybrid beamforming for massive MIMO: A survey." IEEE Communications Magazine 55, no. 9 (2017): 134-141.

[6] Raghavan, Vasanthan, Sundar Subramanian, Juergen Cezanne, Ashwin Sampath, Ozge Hizir Koymen, and Junyi Li. "Single-user versus multi-user precoding for millimeter wave MIMO systems." IEEE Journal on Selected Areas in Communications 35, no. 6 (2017): 1387-1401.

[7] Venkateswaran, Vijay, Florian Pivit, and Lei Guan. "Hybrid RF and digital beamformer for cellular networks: Algorithms, microwave architectures, and measurements." IEEE Transactions on Microwave Theory and Techniques 64, no. 7 (2016): 2226-2243.

[8] Garcia-Rodriguez, Adrian, Vijay Venkateswaran, Pawel Rulikowski, and Christos Masouros. "Hybrid analogdigital precoding revisited under realistic RF modeling." IEEE Wireless Communications Letters 5, no. 5 (2016): 528-531.

[9] Studer, Christoph, Markus Wenk, and Andreas Burg. "System-level implications of residual transmit-RF impairments in MIMO systems." In Antennas and Propagation (EUCAP), Proceedings of the 5th European Conference on, pp. 2686-2689. IEEE, 2011.

[10] Gimenez, Sonia, Sandra Roger, Paolo Baracca, David Martn-Sacristn, Jose F. Monserrat, Volker Braun, and Hardy Halbauer. "Performance evaluation of analog beamforming with hardware impairments for mmW massive MIMO communication in an urban scenario." Sensors 16, no. 10 (2016): 1555.

[11] Mangraviti, Giovanni, Khaled Khalaf, Qixian Shi, Kristof Vaesen, Davide Guermandi, Vito Giannini, Steven Brebels et al. "13.5 a 4-antenna-path beamforming transceiver for 60ghz multi-gb/s communication in 28nm cmos." In Solid-State Circuits Conference (ISSCC), 2016 IEEE International, pp. 246-247. IEEE, 2016.

[12] Maltsev, A. "Channel Modeling and Characterization-MiWEBA." Deliverable 5.1 EU Contract No. FP7-ICT-608637 (2014).

Gabor Expansion for Simultaneous Wireless Power and Information Transfer (SWIPT): Interference Analysis

Hussein Kassab Jérôme Louveaux
Université catholique de Louvain
ICTEAM
Louvain-la-Neuve, Belgium
hussein.kassab@uclouvain.be jerome.louveaux@uclouvain.be

Abstract

Wireless power transfer (WPT) has recently appealed the research community. One application for such technology is powering up future IoT devices and sensors. This article proposes a setup where both WPT and information transfer (IT) coexist sharing the same time and frequency resources in a flexible manner while keeping the interference at a low level. The interference of WPT on IT is analyzed as a performance metric for the proposed scheme. In particular, the influence of WPT side lobes on the main lobes of the information signal is computed to assess its impact on the performance of IT. The proposed scheme uses a Gabor expansion to create both power and information signals. This structure enables to assign resources for energy and information dynamically on the time-frequency lattice providing a high flexibility in the system. We propose to use the Gaussian basis functions to create the WPT signal since they are well localized both in time and frequency, therefore creating the lowest possible interference on the nearby subcarriers. Hence, any multi-carrier system with different modulation schemes can be used for IT in the allowed time-frequency grid such as CP-OFDM and FBMC-OQAM. The expressions of interference are computed and simulation results show the low level of interference created by the power signal on the information signals which validates the possible coexistence of both systems.

1 Introduction

The research on wireless power transfer (**WPT**) has been motivated by the growing interest in the Internet of things (IoT) electronic devices such as battery-free sensor, passive RF identification (RFID), and machine-to-machine (M2M) systems. These future devices can harvest energy from the nearby electromagnetic sources or from dedicated power heads that are designed to maximize their energy transfer. On the other hand, these devices are envisioned to perform information transfer (**IT**) with heterogeneous needs depending on their application. The demand for WPT and IT gave rise to the idea of coexistence between both systems using the same time and frequency resources allowing for greater flexibility in assigning those resources [1]. However, this coexistence faces several obstacles. One significant challenge is the interference caused from the power source on the information transfer system that would significantly affect its performance. This is particularly critical because the power signal is usually much greater than the information signal. To overcome such problem, a new waveform design is required with smart allocation of resources that can reduce the interference effect.

In this paper, we propose a structure that is based on Gabor expansion, where each waveform is represented as a linear combination of shifted basis functions in time and

frequency creating a regular grid also known as time-frequency lattice. Gaussian prototypes are used to generate the power signal which we call **Multi-Gaussian** waveform since they are characterized by their good localization both in time and frequency. Besides, multi-Gaussian is very similar to the multisine signal and is expected to offer similar performance in terms of energy harvesting [1]. Given this good localization property of the power signal, we expect a small interference on the information signals. Any multi-carrier scheme utilizing the same time-frequency grid can then be used for information signals as long as a small guard time/band is used to separate both signals. In this paper, we consider the cyclic-prefix orthogonal frequency-division multiplexing (**CP-OFMD**) and the offset QAM filter-bank multi-carrier (**OQAM-FBMC**). The interference of the multi-Gaussian waveform on both the CP-OFDM and OQAM-FBMC systems is studied. Most interference computation in the literature are based on a Power Spectral Density (PSD) model [3, 4, 5] that considers the out-of-band radiation of the multi-carrier signal. However, the work of [6, 7] showed that the PSD-based model is not precise and proposed a more accurate approach in calculating the interference. In addition, in our case the interference is not random but its deterministic and the PSD-based model is not applicable. In this work, we derive the expressions for the interference of the IT on both aforementioned information systems. As in [8], the receiver architecture (i.e. the filtering operations and processing done at the receiver) is taken into consideration while calculating the interference expressions. Then, the expressions are simulated and the results are discussed in a practical scenario. This scenario shows that the multi-Gaussian power signal has low interference on its neighboring subcarriers and can coexist with both systems under certain conditions.

The paper is organized as follow: First, the proposed scheme and system models are demonstrated in Section 2. The expressions of the interference caused by multi-Gaussian signal on CP-OFDM and OQAM-FBMC in Section 3. Simulated results of the interference is shown in Section 4. Practical Example is presented in Section 5. Finally, the paper is concluded in Section 6.

Notations: scalars are noted as x, vectors are bold as $\mathbf{x}$, x^* is the complex conjugate of x, and sets are represented by a calligraphy letters such as $\mathcal{X}$. n is the discrete index for symbols, m indexes for subcarriers and t is for the continuous time. $*$ is the convolution operation, $\mathbf{E}_x\{y\}$ is the mathematical expectation of y with regards to the random variable x and $\Re\{x\}$ is the real part of x.

2 Proposed Scheme and System models

2.1 Gabor Expansion for coexistence

In this paper, we propose a time-frequency lattice, which is called Gabor system, where different systems with different multi-carrier schemes can coexist together. All systems in this structure share the same time separation T (the symbol duration) and same frequency separation $\Delta f = 1/T$. The proposed system provide flexibility in allocating different multi-carrier systems. However, we must avoid interference caused by frequency side lobes and time tails that would affect the performances of the other systems.

As a result, guard bands and guard periods are proposed to minimize these interference among different systems. One particular example of this proposed structure is illustrated in Fig.1 which shows flexibility of allocating resources for different multi-carrier systems within a certain time-frequency space. It is important to note that all these systems must be synchronized in both time and frequency.

We proposed also to utilize the Gaussian basis function to generate the power signal. The reasons for that are as follows:

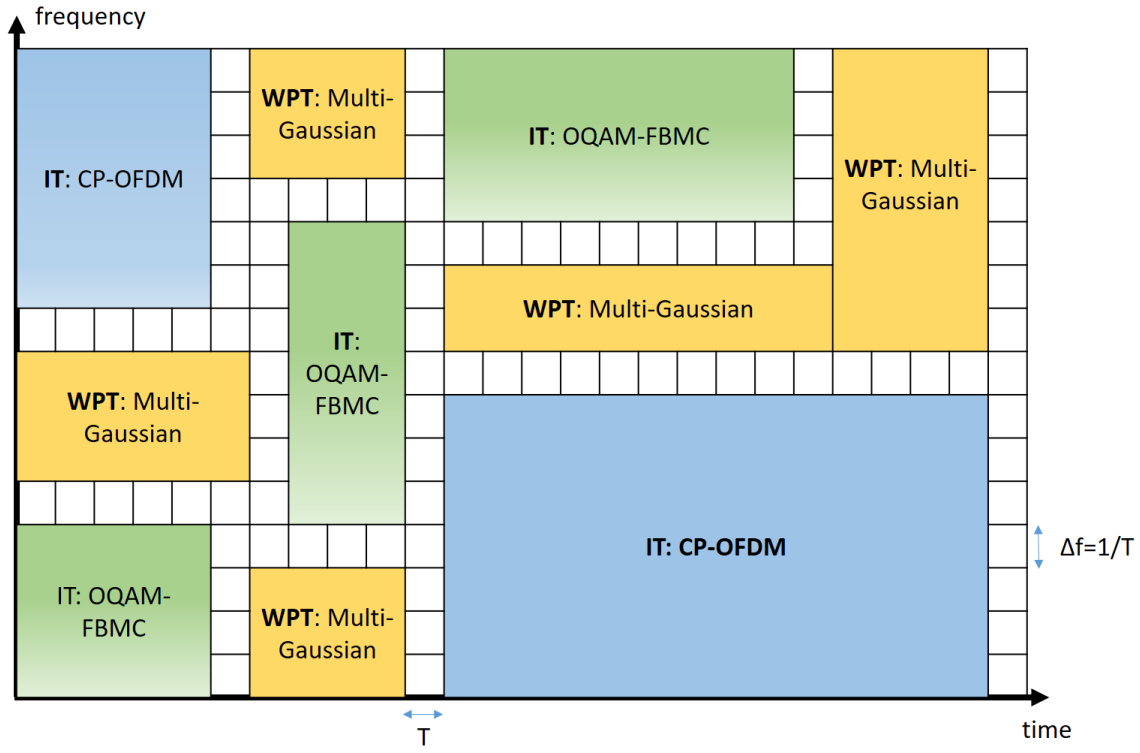


Figure 1: Example of the Proposed Structure of coexistence among different systems

- ◇ There will always be a trade-off between side lobes and time tails. Gaussian functions are well localized in both time and frequency i.e. it has very low side lobes and time tails.
- ◇ It is expected to have good performance in terms of energy harvesting similar to that of the multisine waveform[1].

In our work, we are interested in studying the interference caused by the side lobes only (not the time tails). Therefore we assume the setup illustrated in Fig.2. The information system (whether CP-OFDM or OQAM-FBMC) and the power system share the same bandwidth $\mathcal{B}$. IT occupy a sub-band that belong to a set of information active subcarriers $\mathcal{M}_i$. Whereas, WPT occupy another sub-band that belong to a set of power active subcarriers $\mathcal{M}_p$. In our analysis, we assume we have an additive white Gaussian noise (AWGN) channel and we ignored the path-loss and shadowing of the channel, as well as any propagation delay.

2.2 Power Signal Model: Multi-Gaussian waveform

We consider that a power signal is generated using a Gaussian prototype filter using $\mathcal{M}_p$ active subcarriers out of $\mathcal{M}$ in a fixed bandwidth $\mathcal{B}$. A certain power and phase coefficient $\mathbf{d}_{m_p}[n_p]$ (n_p is the symbol time index) is allocated to a each active subcarrier m_p ($m_p \in \mathcal{M}_p$) during the symbol duration T. The normalized Gaussian prototype filter used is defined as:

$$g(t) = \frac{(2\alpha)^{1/4}}{\sqrt{T}} e^{-\pi\alpha(\frac{t}{T})^2}, \tag{1}$$

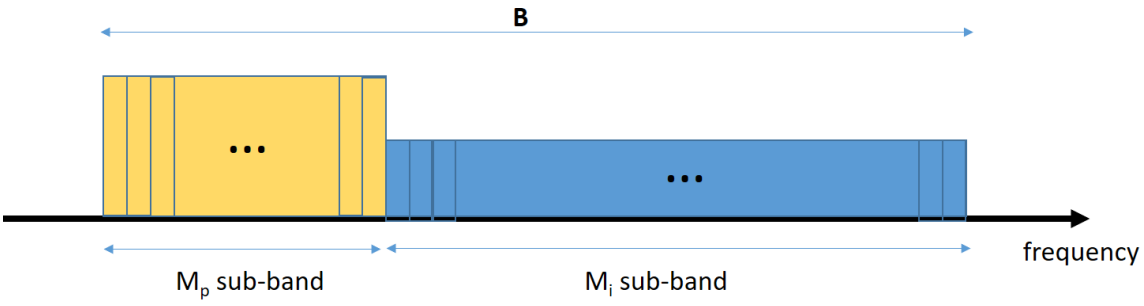


Figure 2: Bandwidth sharing among IT and WPT systems used

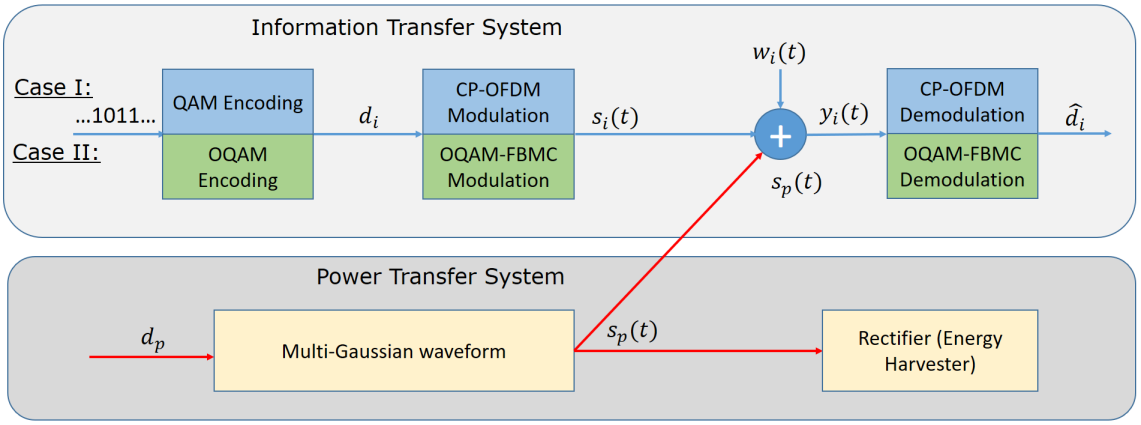


Figure 3: Block diagram representing the IT and WPT systems. Case I: CP-OFDM and Case II: OQAM-FBMC. WPT transmitter is interfering on IT receiver as shown.

where $\alpha > 0$ is the localization parameter. We choose $\alpha = 1$ to have the best localization in both time and frequency. Using the general multi-carrier model in [2] to generate the multi-Gaussian signal, the time-domain signal transmitted on each active subcarrier $m_p \in \mathcal{M}_p$ is expressed as:

$$s_{m_p}(t) = \sum_{n_p \in \mathcal{Z}} \mathbf{d}_{m_p}[n_p] g(t - n_p T) e^{j2\pi m_p \frac{t}{T}}. \tag{2}$$

Hence, the total transmitted signal from the power source (i.e. Multi-Gaussian Waveform) is expressed as:

$$s_p(t) = \sum_{m_p \in \mathcal{M}_p} s_{m_p}(t) = \sum_{m_p \in \mathcal{M}_p} \sum_{n_p \in \mathcal{Z}} \mathbf{d}_{m_p}[n_p] g(t - n_p T) e^{j2\pi m_p \frac{t}{T}}. \tag{3}$$

2.3 Information Signal Model: CP-OFDM

Referring to case I of Fig.3, we consider that the information system that is being influenced by the power signal is a CP-OFDM system. This system has a symbol duration T, cyclic prefix duration T_{CP}, and $\mathcal{M}_i$ active subcarriers out of $\mathcal{M}$ in a fixed bandwidth $\mathcal{B}$. As mentioned above, we assume perfect channel with AWGN. Therefore, the signal at the input antenna of the CP-OFDM receiver can be expressed as:

$$y_i(t) = s_i(t) + s_p(t) + w_i(t), \tag{4}$$

where $s_p(t)$ is the interfering power signal and $w_i(t)$ is the AWGN. At the receiver, assuming perfect synchronization, the cyclic prefix is removed from the signal where only a time window ($f(t) = \frac{1}{\sqrt{T}}$ for $t \in [-\frac{T}{2}, \frac{T}{2}]$) is considered. Then, the Fast Fourier Transform (FFT) is applied to the received windowed signal. The demodulated n_i^{th} symbol on the m_i^{th} subcarrier of the receiver can be expressed as:

$$\hat{\mathbf{d}}_{m_i}[n_i] = \mathbf{d}_{m_i}[n_i] + \sum_{m_p \in \mathcal{M}_p} \eta_{m_p \to m_i}[n_i] + \mathbf{w}_i[n_i], \tag{5}$$

where $\mathbf{w}_i[n_i]$ is the filtered Gaussian noise and $\eta_{m_p \to m_i}[n_i]$ is the interference caused by the m_p^{th} subcarrier of the power source onto the m_i^{th} subcarrier of the information signal which can be expressed as:

$$\eta_{m_p \to m_i}[n_i] = \int_{-\infty}^{\infty} s_{m_p}(t) f(t - n_i(T + T_{CP})) e^{-\frac{j2\pi m_i}{T}(t - n_i T_{CP})} dt. \tag{6}$$

2.4 Information Signal Model: OQAM-FBMC

Referring to case II of Fig.3, we consider that the information system to be OQAM-FBMC with PHYDYAS normalized prototype filter [9] which is expressed as:

$$p(t) = \sqrt{\frac{2}{T}} \sum_{k=-K+1}^{K-1} \frac{G_{|k|}}{K} e^{j2\pi \frac{kt}{KT}}, \quad t \in [-\frac{KT}{2}, \frac{KT}{2}], \tag{7}$$

where $K = 4$ is the overlap factor, and $G_0 = 1$, $G_1 = 0.971960$, $G_2 = \frac{1}{\sqrt{2}}$, and $G_3 = 0.235147$. It is noted also that the filter $p(t)$ is real and symmetric (i.e. $p^*(-t) = p(t)$). The received signal in this case is modulated around the subcarrier of choice then filtered by the prototype filter $p(t)$ where only the real part of the signal is taken. Besides, the received signal is also multiplied by phase factors $\theta_{m_i}[n_i] = j^{(n_i + m_i)}$ and $(-1)^{n_i m_i}$ [9]. The received signal at the OQAM-FBMC receiver can be expressed as equation (4) above. After demodulation, the expression is also similar to equation (5) above. In this case, the interference caused by the m_p^{th} subcarrier of the power source onto the m_i^{th} subcarrier of the OQAM-FBMC signal which can be expressed as:

$$\eta_{m_p \to m_i}[n_i] = \int_{-\infty}^{\infty} s_{m_p}(t)\ p(t - n_i \frac{T}{2})\ (-1)^{n_i m_i}\ j^{(n_i + m_i)}\ e^{-\frac{j2\pi m_i}{T} t} dt. \tag{8}$$

3 Interference Expressions

The total interference inserted from all the active subcarriers of the power source onto the m_i^{th} information subcarrier of the n_i^{th} information symbol can be expressed as follows:

$$\eta_{P \to m_i}[n_i] = \sum_{m_p \in \mathcal{M}_p} \eta_{m_p \to m_i}[n_i]. \tag{9}$$

Then, the interference power caused by all active subcarriers of the power source can be expressed as:

$$I_{P \to m_i} = |\eta_{P \to m_i}[n_i]|^2. \tag{10}$$

In order to ensure fairness in comparison between the two cases (CP-OFDM and OQAM-FBMC), all filters ($f(t)$, $p(t)$, and $g(t)$) are normalized such that both systems transmit with the same energy per symbol. Besides, we assume uniform distribution of symbols $\mathbf{d}_{m_i}[n_i]$ among transmitted active power subcarriers.

3.1 Case I: CP-OFDM

We plug in the expression of the transmitted power signal of one subcarrier shown in (2) into the interference expression in (6). We use the value of the window filter $f(t)$ of the CP-OFDM receiver mentioned in section 2.3 above. Besides, we substitute the value of the Gaussian prototype filter in (1) with $\alpha = 1$. Then, from (10) we get the expression for the interference power as follows:

$$I_{P\rightarrow m_i}^{OFDM} = \frac{\sqrt{2}}{M_p T}\left| \sum_{l\in\{\mathcal{M}_p - m_i\}} \sum_{\tau=\frac{-KT}{2}}^{\frac{KT}{2}} \int_{\frac{-T}{2}}^{\frac{T}{2}} e^{-\pi(\frac{t-\tau}{T})^2 + \frac{j2\pi l}{T}t} dt \right|^2, \tag{11}$$

where τ is the relative time distance between all interfering power symbols onto a given information symbol, and $l = m_p - m_i$ is the spectral distance between a certain active power subcarrier m_p and a given active information subcarrier m_i.

3.2 Case II: OQAM-FBMC

We substitute the expression of the transmitted power signal of one subcarrier shown in (2) into the interference expression in (8). We use the value of the normalized PHYDYAS prototype filter of the OQAM-OFDM receiver shown in (7) above. We plug-in also the Gaussian prototype filter in (1) with $\alpha = 1$. Then, from (10) we get the expression for the interference power as follows:

$$I_{P\rightarrow m_i}^{FBMC} = \frac{1}{\sqrt{2} M_p T K^2}\left| \sum_{l\in\{\mathcal{M}_p - m_i\}} \sum_{\tau=(-K+1)T}^{(K-1)T} \sum_{k=-K+1}^{k=K-1} G_{|k|}\, \Re\left\{ \int_{\frac{-KT}{2}}^{\frac{KT}{2}} e^{-\pi(\frac{t-\tau}{T})^2 + \frac{j2\pi}{T}t(\frac{k}{K}+l)} dt \right\} \right|^2, \tag{12}$$

where τ and $l = m_p - m_i$ are as defined in (11). We should note that a factor of $\frac{1}{2}$ is considered since we are only taking the real part of the signal at the OQAM-FBMC receiver.

4 Simulation results

We consider two systems: the information transfer system (CP-OFDM and OQAM-FBMC) and the power transfer system (multi-Gaussian). Both systems share the same bandwidth $\mathcal{B}$ with total number of subcarriers equals to 128 ($\mathcal{M} = 128$).The WPT system uses 30 active subcarriers ($\mathcal{M}_p = 30$) such that $m_p = [1 \rightarrow 30]$. The IT system uses 98 active subcarriers ($\mathcal{M}_i = 98$) such that $m_i = [31 \rightarrow 128]$.

Given this setup, the interference expressions (11) and (12) represented above were plotted as shown in Fig.4. The following conclusions can be extracted from these results:

- ◇ Generally, the interference is higher for the CP-OFDM as compared to OQAM-FBMC which could be predicted due to the higher side lobes of CP-OFDM.
- ◇ Due to the usage of multi-Gaussian power source, there is a sharp decrease in the interference of the neighboring subcarriers close to the active power subcarriers. This is due to the good localization of the Gaussian filter.
- ◇ The interference become flat when the spectral distance from the power subcarriers is greater than 3.
- ◇ Only one guard band can be considered and the information signal can utilize subcarriers starting from spectral distance equals to 2. That is due to the sharp decrease in interference.

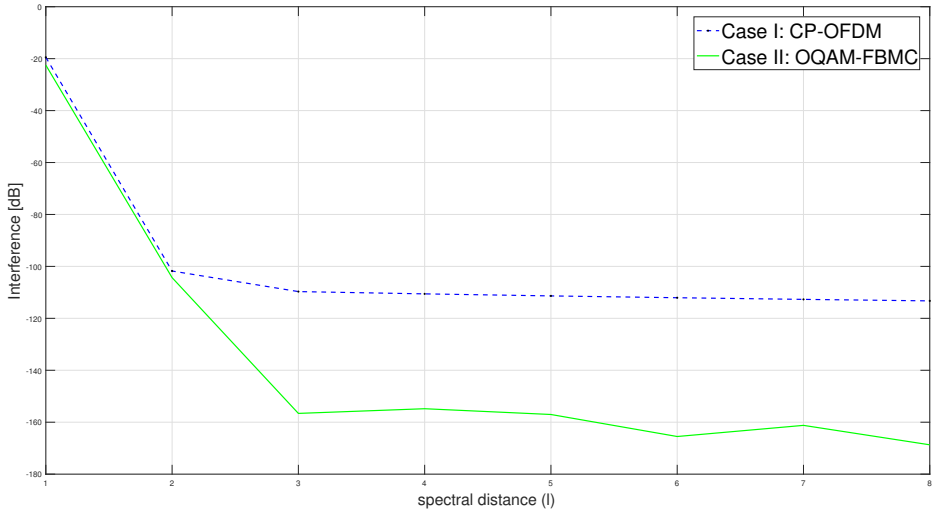


Figure 4: Interference curves for both CP-OFDM and OQAM-FBMC cases

5 Practical Example for coexistence

The simulation results show that the interference values are very low due to the good localization nature of the Gaussian filter used to generate the power signal. This helps in the coexistence of both systems: IT and WPT. A practical example is illustrated in this section to evaluate the value of interference caused by a power transmitter.

Consider a battery-free sensor that requires WPT and IT to operate (i.e. SWIPT device). We consider that both systems transmit at $f_c = 2.4GHz$ and they share the same bandwidth $\mathcal{B} = 1.92MHz$ with $\mathcal{M} = 128$ subcarriers and frequency separation $\Delta f = 15kHz$. The information source transmits at a power of $P_{I,tx} = 0dBm$ while the power source transmits at a power of $P_{P,tx} = 23dBm$. The sensor is present in a fixed place where it is located at a near distance from the power head ($d_p = 2m$) while its further away from the information transmitter ($d_i = 35m$) which potentially create a gap between the power levels of both systems. Both signals will experience a path-loss components before reaching the information receiver. We use the free space path-loss model to get: $L_p = 46.0726dB$ and $L_i = 70.9334dB$ for the power and information signals respectively. We will ignore the channel shadowing and scattering effect for both signals in this example. Hence, the received power at the antenna of the information signal and the power signal are $P_I = -70.9334dBm$ and $P_P = -23.0726dBm$ respectively.

To evaluate the performance of the information subcarrier m_i, we will use the signal-to-interference-ration (SINR) as a performance metric which is defined as:

$$SINR(m_i) = \frac{P_{m_i,I}}{I_{P\rightarrow m_i} + N}, \tag{13}$$

where $P_{m_i,I}$ is the power allocated for subcarrier m_i of the information signal at the antenna receiver, $I_{P\rightarrow m_i}$ is the interference caused by the power signal on the information signal, and N is the power of the noise experienced by the information signal ($N = -141dB$).

As an example, we consider the adjacent information subcarrier $m_i = 31$ to be the guard band and we want to evaluate the SINR at $m_i = 32$. Assuming uniform distribution of power among information subcarriers (98 subcarriers), hence $P_{31,I} =$

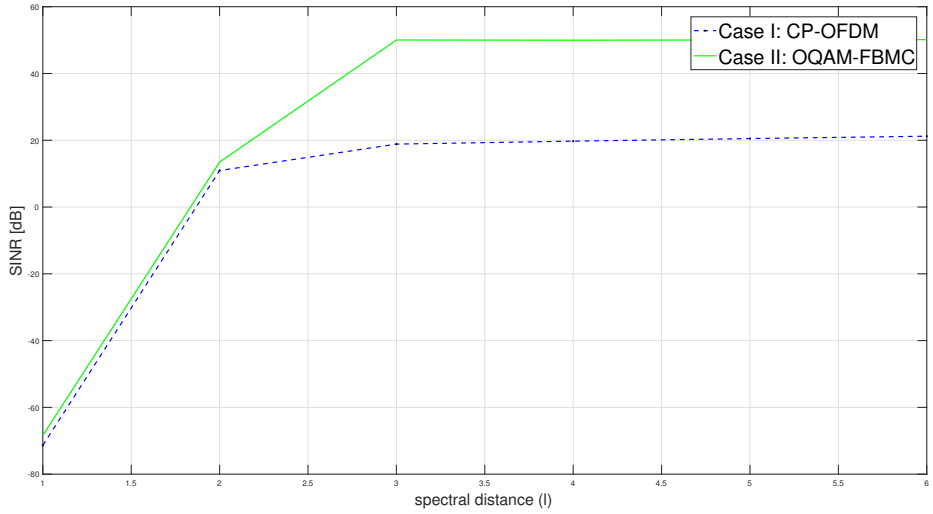


Figure 5: SINR for information subcarriers of Case I: CP-OFDM and Case II: OQAM-FBMC

$-90.8456dBm$. Then, we evaluate the interference caused by all active subcarriers of the power signal on this particular information subcarrier from the interference curve obtained in the simulation section above ($I^{OFDM}_{P\to m_i} = -101.7777dBm$ and $I^{FBMC}_{P\to m_i} = -109.7005dBm$). Then, we calculate the SINR values for both IT cases using (13) as follows:

$$SINR^{OFDM}(31) = \frac{P_{31,I}}{I_{P\to 31} + N} = 10.9316dB, \tag{14}$$

$$SINR^{FBMC}(31) = \frac{P_{m_i,I}}{I_{P\to 31} + N} = 13.4862dB. \tag{15}$$

We followed the same steps to obtain the SINR curves for all other information subcarriers as shown in Fig.5. Based on the results, we observe with this practical example that the SINR of OQAM-FBMC is better than CP-OFDM. Also, one guard band only is recommended after which the information subcarriers can utilize the remaining subcarriers. SINR also becomes flat after $l = 3$.

6 Conclusion

In this work, we provided a structure where information and power signals can coexist in flexible manner sharing the same bandwidth. The power signal is proposed to be generated by a Gaussian function that is well localized in time and frequency and that is expected to have good performance in terms of energy harvesting since its similar to the multisine. The interference of the power signal is studied on two information multi-carrier systems CP-OFDM and OQAM-FBMC. Results shows low level of interference injected from the power signal onto the both cases of information signals which means that IT and WPT can coexist by considering only one guard band.

References

[1] B. Clerckx and E. Bayguzina, "Waveform design for wireless power transfer," IEEE Trans. Signal Process., vol. 64, no. 23, pp. 63136328, Dec. 2016

[2] A. S. ahin, I. Guvenc, and H. Arslan, "A survey on multi-carrier communications: Prototype filters, lattice structures, and implementation aspects," IEEE Commun. Surveys Tutorials, vol. 16, no. 3, pp. 13121338, Aug. 2014.

[3] LG. Baltar, DS. Waldhauser, JA. Nossek, "Out-of-band radiation in multi-carrier systems: a comparison," in multi-carrier Spread Spectrum, page numbers, vol. 1 (Springer Netherlands, 2007), pp. 107–116.

[4] H. Zhang, DL. Ruyet, M. Terr, "Spectral efficiency comparison between OFDM/OQAM and OFDM based CR networks," Wireless Commun. Mobile Comput. Wiley. 9, 1487–1501 (2009)

[5] T. Weiss, J. Hillenbrand, A. Krohn, FK. Jondral, "Mutual interference in OFDM-based spectrum pooling systems," in IEEE 59th Vehicular Technology Conference, vol. 4 (IEEE Piscataway, 2004), pp. 1873–1877

[6] Q. Bodinier, F. Bader, J. Palicot, "Modeling Interference Between OFDM / OQAM and CP-OFDM: Limitations of the PSD-Based Model," in International Conference on Telecommunications (ICT), Thessaloniki, 2016.

[7] Y. Medjahdi, M. Terre, D. L. Ruyet, and D. Roviras, "Interference tables: a useful model for interference analysis in asynchronous multi-carrier transmission," EURASIP Journal on Advances in Signal Processing, vol. 2014, no. 54, pp. 117, 2014.

[8] Q. Bodinier, F. Bader, and J. Palicot, "Coexistence in 5G: Analysis of cross-interference between OFDM/OQAM and legacy users," in Proc. IEEE Globecom Workshops (GC Wkshps), Dec. 2016, pp. 16.

[9] M. Bellanger, "Specification and Design of a Prototype Filter for Filter Bank Based Multi-carrier Transmission," in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings, vol. 4, 2001, pp. 2417–2420.

A Novel Low-Complexity Robust Distributed Beamformer

Andreas I. Koutrouvelis, Thomas W. Sherson, Richard Heusdens and Richard C. Hendriks
Delft University of Technology

We propose a new low-complexity robust linearly constrained beamformer which utilizes a set of linear equality constraints to reduce the cross power spectral density matrix to a block-diagonal form. The proposed beamformer is robust to relative acoustic transfer function (RATF) estimation errors and to target activity detection (TAD) errors. Two variants of the proposed beamformer are presented and evaluated in the context of multi-microphone speech enhancement, and are compared with other state-of-the-art low-complexity beamformers in terms of robustness to RATF estimation errors and TAD errors.

All compared linearly constrained beamformers are special cases of the following general linearly constrained quadratic minimization problem

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{P} \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{\Lambda} = \mathbf{f}^H, \quad \mathbf{\Lambda} = \begin{bmatrix} \mathbf{a} & \mathbf{b}_1 & \cdots & \mathbf{b}_r \end{bmatrix}, \text{ and } \mathbf{f} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^H,$$

where $\mathbf{P}$ is a cross-power spectral density matrix which is different for each method as shown in the following Table.

Method	$\mathbf{P}$	Constraints	Target activity detection
MPDR	$\mathbf{P_y}$	$\mathbf{w}^H\mathbf{a} = 1$	no
MVDR	$\mathbf{P_n}$	$\mathbf{w}^H\mathbf{a} = 1$	yes
DS	$\mathbf{I}$	$\mathbf{w}^H\mathbf{a} = 1$	no
LCMP	$\mathbf{P_y}$	$\mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H$	no
LCMV	$\mathbf{P_n}$	$\mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H$	yes
LCDS	$\mathbf{I}$	$\mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H$	no
BDLCMP (proposed)	$\bar{\mathbf{P}}_\mathbf{y}$	$\mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H$	no
BDLCMV (proposed)	$\bar{\mathbf{P}}_\mathbf{n}$	$\mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H$	yes

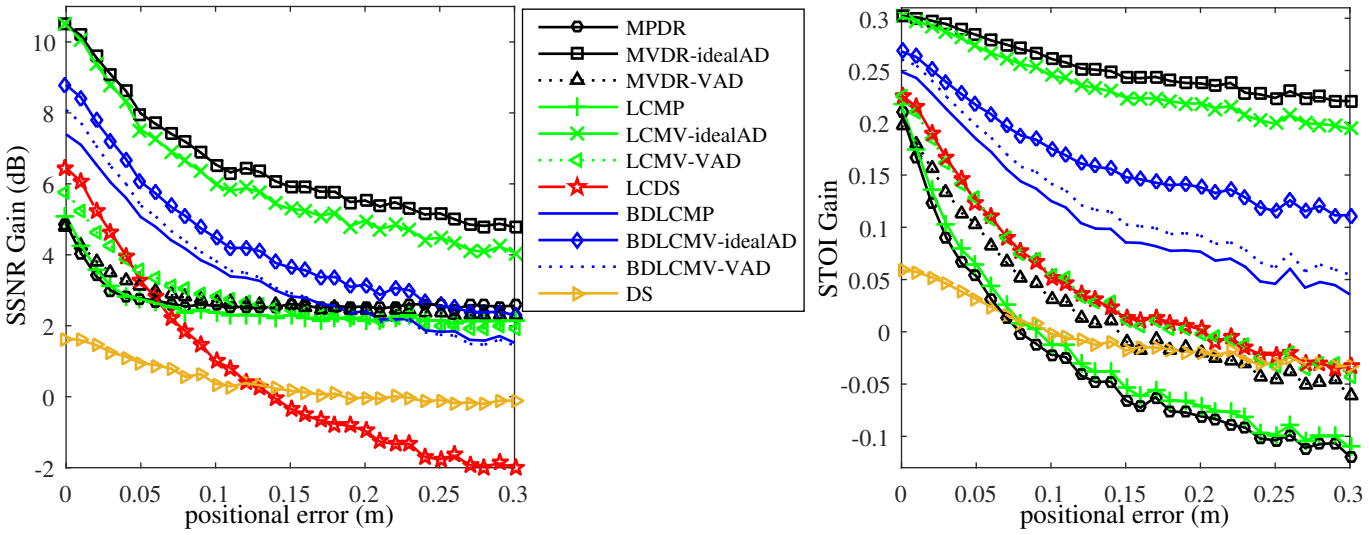


Figure 1: Comparison of the beamformers in the above Table as a function of positional error. The methods that depend on a TAD are computed using an ideal TAD and a state-of-the-art voice activity detector (VAD).

Zero Secrecy Leakage for Multiple Enrollments of Physical Unclonable Functions

Lieneke Kusters* c.j.kusters@tue.nl
Onur Günlü** onur.gunlu@tum.de
Frans M.J. Willems* f.m.j.willems@tue.nl

*Eindhoven University of Technology
Eindhoven, The Netherlands

**Technical University of Munich
Munich, Germany

Abstract

We use physical unclonable functions (PUFs) to generate secret keys. We analyze the performance of the helper data scheme when the enrollment process is repeated multiple times. We show that codes exist such that the scheme remains secure after two enrollments, when all PUF observations are performed over the same channel. Furthermore, we show that a fuzzy commitment scheme remains secure after any number of enrollments, for PUF sources that meet a certain symmetry condition. We show that the temperature dependent model for SRAM-PUF meets this symmetry condition. Furthermore, we argue that many source-channel model pairs exist that meet the symmetry condition, and give some examples. *

1 Introduction

Sensitive data are protected by using secret keys. We use physical unclonable functions (PUFs) to construct such secret keys. A PUF corresponds to a response of a physical device to a challenge. This response is device-specific, reliable, and unpredictable. An attacker may try to guess the key, however, he does not have access to the PUF response. We use the PUF response to generate a random key that is unpredictable for the attacker. Furthermore, we reconstruct the same key at any time by using another response of the same PUF. Two responses of the same PUF device are similar, but not exactly the same due to noise. Therefore, we use an error-correcting code and a helper data scheme to ensure that exactly the same key can be reproduced from a noisy response of the same PUF; see Fig. 1 for a helper data scheme.

1.1 Helper Data Scheme

In a helper data scheme, as depicted in Fig. 1, we distinguish two phases: an enrollment and a reconstruction phase. During the enrollment phase, a key s is generated

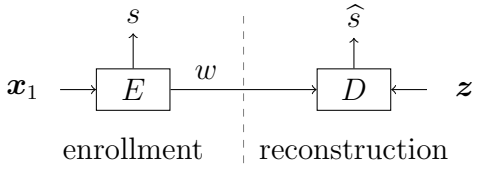


Figure 1: Helper data scheme.

*This work was funded by Eurostars-2 joint programme with co-funding from the EU Horizon 2020 programme under the E! 11897 RESCURE project.

for the first time, based on a response $\boldsymbol{x}$ of the PUF. In addition to the key, a helper message w is generated. The helper message w provides sufficient information to reconstruct the same key s from another response $\boldsymbol{z} \approx \boldsymbol{x}$ of the same PUF during the reconstruction phase. An attacker cannot observe the PUF responses $\boldsymbol{x}$ and $\boldsymbol{z}$, but the helper message is communicated over a public channel, and it may be observed by the attacker. Therefore, the helper message w should not reveal information about the s to an attacker.

The code used in the helper data scheme produces a secret $s \in \mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$ and a helper message w, based on a PUF response $\boldsymbol{x}$ of length $n > 0$, such that the following conditions are satisfied

$$\Pr(\widehat{S} \neq S) \leq \delta, \tag{1}$$

$$\frac{1}{n} H(S) + \delta \geq \frac{1}{n} \log_2 |\mathcal{S}| \geq R_s - \delta, \tag{2}$$

$$\frac{1}{n} I(W; S) \leq \delta \tag{3}$$

for a $\delta > 0$ and n large enough. Here, R_s is an achievable secret-key rate, and the maximum achievable secret-key rate, i.e., secret-key capacity, is known to be $C_s = I(\boldsymbol{X}_1; \boldsymbol{Z})$ [1, 2].

1.2 Literature Review and Main Contributions

In the key agreement literature, only a single enrollment is assumed to be performed for each PUF device. We are interested in a situation that the enrollment procedure is repeated multiple times. This may happen in practice, when, for example, the key is replaced with a new one or the overlying protocol that includes the first enrollment is repeated for some reason. Note that the generated keys will not (necessarily) be independent. We assume that during each PUF enrollment, the previous key is replaced with a new key and a corresponding helper message is published. The decoder uses the most recent helper message to reconstruct the corresponding key. The previous helper messages may all be used by an attacker to derive information about the key.

We have discussed multiple enrollment of an SRAM-PUF with the fuzzy commitment scheme [3] in [4], and showed that for a given model of the SRAM-PUF, the fuzzy commitment scheme remained secure, i.e., (3) is satisfied, also when the multiple enrollments were performed. We extended the proof to the syndrome method in [5].

In the current work, we generalize our results. We show that in the case of two enrollments when all PUF measurements are done via the same channel, there exists a code that satisfies (1)-(3) and achieves the secret-key capacity. Furthermore, we show that the fuzzy commitment scheme remains secure for any number of enrollments given that the PUF response meets a certain symmetry condition. Finally, we extend the model that we used in [4, 5] to a temperature-dependent SRAM-PUF model, similar to ring oscillator PUF models we had in [6], and show that it meets the symmetry condition.

1.3 Notation and PUF Response Assumptions

We use upper case letters to denote random variables and lower case letters to denote their realizations. The symbol $\boldsymbol{x}$ refers to a PUF response used for enrollment and $\boldsymbol{z}$ a PUF response used for reconstruction. We use different symbols to make it easier to follow our reasoning in equations; however, the statistical properties of both responses are the same. Assume that the PUF responses are binary vectors of length n; thus, $\boldsymbol{x} \in \{0, 1\}^n$ and $\boldsymbol{z} \in \{0, 1\}^n$. Furthermore, suppose the values in the vector are independent

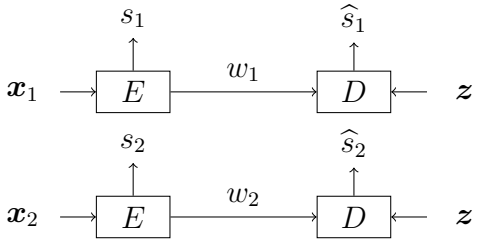


Figure 2: Two enrollments scheme.

identically distributed (i.i.d.), i.e. $\Pr(\boldsymbol{X} = \boldsymbol{x}) = \prod_{i=1}^{n} \Pr(X(i) = x(i))$. Assume that multiple responses $(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots)$ are observed, and the probability distributions are time and permutation invariant, e.g., $\Pr(\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2) = \Pr(\boldsymbol{X}_1 = \boldsymbol{x}_2, \boldsymbol{X}_2 = \boldsymbol{x}_1)$. The function $H(X)$ is the entropy function for a random variable. It follows from our assumptions above that

$$H(\boldsymbol{X}_j) = H(\boldsymbol{Z}) = nH(X_1) \qquad \forall j \in \{1, 2, \dots\}, \tag{4}$$

$$H(\boldsymbol{X}_i \boldsymbol{X}_j) = nH(X_1 X_2) \qquad i \neq j,\ \forall i, j \in \{1, 2, \dots\}, \tag{5}$$

$$H(\boldsymbol{X}_i \boldsymbol{X}_j \boldsymbol{X}_k) = nH(X_1 X_2 X_3) \qquad i \neq j \neq k,\ \forall i, j, k \in \{1, 2, \dots\}. \tag{6}$$

2 Multiple Enrollments

We are interested in the scenario where multiple keys and helper messages are generated independently from different measurements of the same PUF. As an example, we show a two-enrollment scenario in Fig. 2. During each enrollment $j \in \{1, 2\}$, a key s_j and corresponding helper message w_j are generated based on an observation $\boldsymbol{x}_j$ of the PUF. The helper message w_j should contain sufficient information such that a decoder can reconstruct the secret s_j when an additional observation $\boldsymbol{z}$ of the PUF is available. An attacker given all helper messages $(w_1, w_2, \dots)$; however, should not learn any information about any of the secrets s_j. We assume that the same code is used by all encoders and decoders. It is required that each key is uniformly distributed. Note that it is not required that the keys are independent. It is straightforward to show that secrecy is leaked about all secret keys when a single key is compromised.

Each encoder generates a key $s_j \in \mathcal{S} = \{1, 2, \dots, |\mathcal{S}|\}$, and a corresponding helper message $w_j \in \mathcal{W} = \{1, 2, \dots, |\mathcal{W}|\}$ by using the same code. We are interested in finding codes that achieve the following conditions for $l \geq 1$ enrollments:

$$\Pr(\widehat{S}_1 \neq S_1 \cup \widehat{S}_2 \neq S_2 \cup \cdots \cup \widehat{S}_l \neq S_l) \leq \delta, \tag{7}$$

$$\frac{1}{n} H(S_t) + \delta \geq \frac{1}{n} \log_2 |\mathcal{S}| \geq R_s - \delta \qquad \forall t \in \{1, 2, \dots, l\}, \tag{8}$$

$$\frac{1}{n} I(W_1 W_2 \dots W_l; S_t) \leq \delta \qquad \forall t \in \{1, 2, \dots, l\} \tag{9}$$

for a $\delta > 0$ and n large enough. Here, R_s is an achievable secret-key rate, and we are interested in finding the set of achievable secret-key rates that satisfies (7)-(9) for any number of enrollments.

We can find a straightforward upper bound on the achievable rates. Since any enrollment follows exactly the same procedure and is also based on a response with the same statistical properties as the first enrollment, it should be clear that we cannot achieve a higher rate than the rate achieved for a single enrollment. We prove this upper bound in Appendix A.

2.1 Two Enrollments

For the two enrollments scenario shown in Fig. 2, we have the following result.

Theorem 1. *The secret-key capacity for each secret key in a two-enrollment setup defined above is $I(X_1;Z) = I(X_2;Z)$, which is equal to the secret-key capacity for a single enrollment setup.*

Proof. First, we define a random labeling $G : \{0,1\}^n \to (\mathcal{S}, \mathcal{W})$, with $\mathcal{S} = \{1, 2, \ldots, 2^{nR_S}\}$ and $\mathcal{W} = \{1, 2, \ldots, 2^{nR_W}\}$. An encoder uses the labeling G to map an observation vector $\boldsymbol{x}$ to a corresponding secret key and helper message pair (s, w). We distinguish two types of decoders: a regular decoder and a virtual decoder. The regular decoder performs the reconstruction of the secret on the decoding side of the helper data scheme. The virtual decoder is an imaginary decoder which we define to help us prove that the uniformity and leakage conditions are satisfied. Both decoders are based on joint typicality decoding, where we define the set of weakly typical sequences $\mathcal{A}_\epsilon^n(X)$ with respect to P_X as in [7].

The regular decoder reconstructs an observation vector $\boldsymbol{x}_i$ by using his observation $\boldsymbol{z}$ and the helper message w_i. That is, it decodes $\widehat{\boldsymbol{x}}_i$ when the labeling $G(\widehat{\boldsymbol{x}}_i) = (*, w_i)$, where $*$ corresponds to any value that is in $\mathcal{S}$, and $(\widehat{\boldsymbol{x}}_i, \boldsymbol{z}) \in \mathcal{A}_\epsilon^n(X_i Z)$. When the decoder has successfully decoded $\boldsymbol{x}$, it can also reconstruct the secret s by using the labeling G.

The virtual decoder decodes both observations $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ from all helper messages (w_1, w_2) and one of the secrets s_i. First, it decodes $\widetilde{\boldsymbol{x}}_i$, such that the labeling $G(\widetilde{\boldsymbol{x}}_i) = (s_i, w_i)$, and $(\widetilde{\boldsymbol{x}}_i) \in \mathcal{A}_\epsilon^n(X_i)$. Then, it can decode the other observation $\boldsymbol{x}_j$, using the corresponding helper message w_j and the already decoded observation $\boldsymbol{x}_i$, with joint typicality decoding, as with the regular decoder. That is, it decodes $\widetilde{\boldsymbol{x}}_j$ when the labeling $G(\widetilde{\boldsymbol{x}}_j) = (*, w_j)$, and $(\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_j) \in \mathcal{A}_\epsilon^n(X_1 X_2)$.

Choosing $R_W = H(X_1|Z)$ and $R_W + R_S = H(X_1)$, it follows from the Slepian-Wolf Theorem, see, e.g., [7], that error-probabilities can be made arbitrarily small for both decoders by increasing n. Therefore, a code exists such that $R_S = I(X_1;Z) = H(X_1) - H(X_1|Z)$ and $R_W = H(X_1|Z)$ and both the regular and virtual decoders are successful with high probability. We say that the probability of error by the decoders is at most ϵ_n when a code of length n is used.

Given that there exists such a code, we can find an upper bound on the entropy of the helper messages and the secrets as $H(W_t) \leq nR_W = nH(X_1|Z)$ and $H(S_t) \leq nR_S = nI(X_1;Z)$. Now, the secrecy leakage about the secret key s_t via the two helper messages is:

$$\begin{aligned}
&I(S_t; W_1 W_2) = H(S_t) + H(W_1 W_2) - H(S_t W_1 W_2) \\
&\leq n(I(X_1;Z) + 2H(X_1|Z)) - H(S_t W_1 W_2 \boldsymbol{X}_1 \boldsymbol{X}_2) + H(\boldsymbol{X}_1 \boldsymbol{X}_2 | W_1 W_2 S_t) \\
&\overset{(a)}{\leq} n(H(X_1) - H(X_1|Z) + H(X_1|Z) + H(X_2|Z) - H(X_1 X_2)) + 1 + 2n\epsilon_n \\
&\overset{(b)}{=} 1 + 2n\epsilon_n
\end{aligned} \tag{10}$$

where (a) follows from Fano's inequality for the virtual decoder, i.e. $H(\boldsymbol{X}_1 \boldsymbol{X}_2 | W_1 W_2 S_t) \leq 1 + 2n\epsilon_n$ with error probability $\epsilon_n > 0$ and (b) follows because $P_{X_1 X_2} = P_{X_2 Z}$ for the measurement model we consider. Thus, the security condition in (9) is satisfied.

Finally we obtain from Fano's inequality for the virtual decoder that

$$\begin{aligned}
&nH(X_t) = H(\boldsymbol{X}_t) = H(\boldsymbol{X}_t W_t S_t) \\
&\leq H(S_t) + H(W_t) + H(\boldsymbol{X}_t | W_t S_t) \\
&\leq H(S_t) + nH(X_1|Z) + 1 + n\epsilon_n
\end{aligned} \tag{11}$$

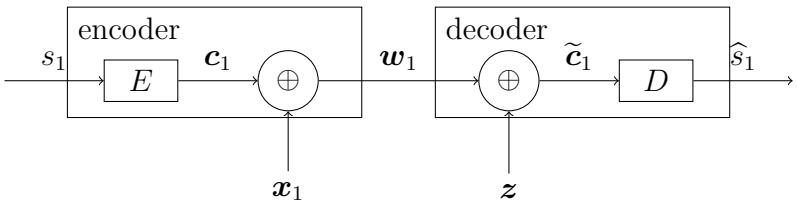


Figure 3: Fuzzy commitment scheme.

and thus $\frac{1}{n}H(S_t) \geq I(X_1;Z) - \epsilon_n - \frac{1}{n}$. Therefore, for any $\delta > 0$ we obtain that $\frac{1}{n}H(S_t) + \delta \geq I(X_1;Z) = \frac{1}{n}\log_2|\mathcal{S}| \geq R_S - \delta = I(X_1;Z) - \delta$, for a large enough n and an ϵ_n that tends to zero as $n \to \infty$. Thus, the uniformity condition for the secret keys holds.

We conclude that a code exists such that all conditions for the two enrollment scenario are met and that achieves secret-key rates

$$R_S = I(X_1;Z) = I(X_2;Z). \tag{12}$$

□

3 Zero Secrecy Leakage for Symmetric PUFs

In the following, we show that zero secrecy leakage occurs for any number of enrollments, when a linear code is used in the fuzzy commitment scheme and the PUF source has a certain type of symmetry.

3.1 Fuzzy Commitment Scheme

First, we introduce the fuzzy commitment scheme for secret-key binding, see Fig. 3. On the encoder side, a secret-key is encoded into a binary codeword that is bound to the PUF output. The helper message $\boldsymbol{w}_1$ is the modulo-2 sum of the codeword $\boldsymbol{c}_1$ and the PUF output $\boldsymbol{x}_1$. A decoder can now reconstruct a noisy version of the codeword $\widetilde{\boldsymbol{c}}_1 = \boldsymbol{w}_1 \oplus \boldsymbol{z} = \boldsymbol{c}_1 \oplus (\boldsymbol{x}_1 \oplus \boldsymbol{z})$ by summing his observation of the PUF and the received helper message modulo-2. As long as there are not too many errors between the two observations $\boldsymbol{x}_1$, and $\boldsymbol{z}$, the decoder can reconstruct the secret from $\widetilde{\boldsymbol{c}}_1$. This procedure is repeated for each PUF enrollment. Furthermore, we use a linear code $E : \{0,1\}^n \to \mathcal{C}$ to encode each secret into a binary codeword as $E(s) = \boldsymbol{c}$. Since the code is linear, it follows that the modulo-2 sum of two codewords is also a codeword $(\boldsymbol{c} \oplus \boldsymbol{c}') \in \mathcal{C}$.

3.2 Symmetry Property for Zero Secrecy Leakage

We list our results from [5] that show that the fuzzy commitment scheme does not leak any information about the secret key, when the PUF observations have the following symmetry property for all $x_1, x_2, \ldots$:

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots) = \Pr(X_1 = \overline{x_1}, X_2 = \overline{x_2}, \ldots) \tag{13}$$

where $\bar{x}$ is the one's complement of x. Firstly, we have

$$\begin{aligned}\Pr(\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, \ldots) &= \Pr(\boldsymbol{X}_1 = \overline{\boldsymbol{x}_1}, X_2 = \overline{\boldsymbol{x}_2}, \ldots) \\ &= \Pr(\boldsymbol{X}_1 = \boldsymbol{x}_1 \oplus \boldsymbol{x}_j, \boldsymbol{X}_2 = \boldsymbol{x}_2 \oplus \boldsymbol{x}_j, \ldots)\end{aligned} \tag{14}$$

where $\boldsymbol{x}_j$ is an observed vector. Then, we can derive that the probability distribution for any set of generated helper messages and given the l^{th} secret s_l and its corresponding codeword $\boldsymbol{c}_l$,

$$\begin{aligned}
&\Pr(\boldsymbol{W}_1 = \boldsymbol{w}_1, \ldots, \boldsymbol{W}_l = \boldsymbol{w}_l | \boldsymbol{C}_l = \boldsymbol{c}_l) \\
&= \sum_{\boldsymbol{c}_1 \in \mathcal{C}} \cdots \sum_{\boldsymbol{c}_{l-1} \in \mathcal{C}} \Pr(\boldsymbol{X}_1 = \boldsymbol{w}_1 \oplus \boldsymbol{c}_1, \ldots, \boldsymbol{X}_l = \boldsymbol{w}_l \oplus \boldsymbol{c}_l) \\
&\overset{(a)}{=} \sum_{\boldsymbol{c}'_1 \in \mathcal{C}} \cdots \sum_{\boldsymbol{c}'_{l-1} \in \mathcal{C}} \Pr(\boldsymbol{X}_1 = \boldsymbol{w}_1 \oplus \boldsymbol{c}'_1, \ldots, \boldsymbol{X}_l = \boldsymbol{w}_l \oplus \boldsymbol{0}) \\
&\overset{(b)}{=} \Pr(\boldsymbol{W}_1 = \boldsymbol{w}_1, \ldots, \boldsymbol{W}_l = \boldsymbol{w}_l), \qquad (15)
\end{aligned}$$

where (a) follows from the linearity of the code, and (b) follows because there is no longer a dependency on the value of c_l. The above derivation can be repeated for any of the secrets s_j, so we have

$$H(\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_l | S_j) = H(\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_l). \qquad (16)$$

We conclude that the leakage about any secret S_j, by all the observation vectors

$$I(\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_l; S_j) = H(\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_l) - H(\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_l | S_j) = 0. \qquad (17)$$

3.3 SRAM-PUF Model Under Varying Ambient Temperature

Suppose the power-on value of an SRAM is used to generate the PUF response used in the fuzzy commitment scheme. We use the statistical model presented in [8] that models the temperature dependent behavior of the SRAM values. Each SRAM cell has two hidden model variables that define the probability that a bit one is observed at an ambient temperature T. The first hidden variable m defines the bias of the cell. The second hidden variable d defines how the one-probability changes with temperature. For an SRAM cell with given hidden variables m and d, the j^{th} observation at temperature $T^{(j)}$ is modeled as

$$r^{(j)}(T^{(j)}) = \begin{cases} 0 & \text{if } m + n^{(j)} + d \cdot T^{(j)} \leq \tau, \\ 1 & \text{if } m + n^{(j)} + d \cdot T^{(j)} > \tau \end{cases} \qquad (18)$$

where $n^{(j)}$ represents the realization of noise sampled in each measurement according to a Gaussian distribution $\mathcal{N}(0, 1)$. The probability that a one is observed at temperature T for this cell, is given by $Q(-m - d \cdot T + \tau)$, where $Q(\cdot)$ is the Q-function. The hidden variables m and d are assumed to be unknown. These cell properties are a result of random variations in the production process and are modeled as independent samples of the random variables, respectively, $M \sim \mathcal{N}(\mu_M, \sigma_M)$ and $D \sim \mathcal{N}(0, \sigma_D)$ for each SRAM cell. Therefore, for a cell with unknown values for the hidden variables, the one-probability at temperature $T^{(j)}$ is

$$\Pr(R^{(j)} = 1) = \int \int Q(-m - d \cdot T^{(j)} + \tau) p_M(m) p_D(d) \, \mathrm{d}m \, \mathrm{d}d. \qquad (19)$$

Assume that the SRAM cells are unbiased (i.e., on average the probability that a one is observed for any SRAM cell, is equal to the probability that a zero is observed),

so $\mu_M = \tau$. For l observations of an SRAM cell, at various *given* temperatures $T^l = (T^{(1)}, T^{(2)}, \ldots, T^{(l)})$, we have

$$\begin{aligned}
&\Pr(R^l = r^l) \\
&= \int\int \prod_{j=1}^{l} Q(-m - d \cdot T^{(j)})^{r^{(j)}} Q(m + d \cdot T^{(j)})^{\overline{r^{(j)}}} p_M(m) p_D(d) \, \mathrm{d}m \, \mathrm{d}d \\
&= \int\int \prod_{j=1}^{l} Q(m + d \cdot T^{(j)})^{r^{(j)}} Q(-m - d \cdot T^{(j)})^{\overline{r^{(j)}}} p_M(-m) p_D(-d) \, \mathrm{d}m \, \mathrm{d}d \\
&\overset{(a)}{=} \int\int \prod_{j=1}^{l} Q(-m - d \cdot T^{(j)})^{\overline{r^{(j)}}} Q(m + d \cdot T^{(j)})^{r^{(j)}} p_M(m) p_D(d) \, \mathrm{d}m \, \mathrm{d}d \\
&= \Pr(R^l = \overline{r^l})
\end{aligned} \tag{20}$$

where (a) follows from the symmetry properties $p_M(m) = p_M(-m)$ and $p_D(d) = p_D(-d)$.

Since the hidden model variables are i.i.d. over all SRAM cells, we have for observation vectors $\boldsymbol{r}^l = (r_1^l, r_2^l, \ldots, r_n^l)$ corresponding to l observations of n SRAM cells at temperatures T^l that

$$\begin{aligned}
\Pr(\boldsymbol{R}^l = \boldsymbol{r}^l) &= \Pr\left((R_1^l, R_2^l, \ldots, R_n^l) = (r_1^l, r_2^l, \ldots, r_n^l)\right) \\
&\overset{(a)}{=} \prod_{i=1}^{n} \Pr(R_i^l = r_i^l) \overset{(b)}{=} \prod_{i=1}^{n} \Pr(R_i^l = \overline{r_i^l}) \\
&= \Pr(\boldsymbol{R}^l = \overline{\boldsymbol{r}^l})
\end{aligned} \tag{21}$$

where (a) follows from independence of the SRAM cells, and (b) follows by (20). Therefore, we conclude that the temperature dependent SRAM-PUF model given in [8] meets the symmetry condition in (13) that results in zero secrecy leakage when the fuzzy commitment scheme is used.

3.4 Other PUF Models with the Symmetry Property

We remark that a PUF response $\boldsymbol{x}$ is a noisy observation of a hidden source through a measurement channel. We can show that there is a big set of source-channel model pairs that satisfy the symmetry property in (13) in addition to the SRAM-PUF model under varying temperature conditions, as discussed in Section 3.3. For instance, any binary-input symmetric memoryless measurement channel (see, e.g., [9] for its definition) such as dependent binary symmetric PUF measurement channels [10] satisfies this equality if the hidden source is symmetric.

We also give an example source-channel model pair where both the source and channel are asymmetric but the outputs are symmetric to further illustrate that the symmetry property in (13) is not limited to a small set of models. Consider an asymmetric hidden source with one-probability $\Pr(Y = 1) = 4/5$, and an asymmetric measurement channel that is given by the Z-channel with parameter $z = 3/8$, see Fig. 4. Now single channel observations are symmetric, that is $\Pr(X = 1) = \Pr(X = 0) = 1/2$. Furthermore, for two observations $P_{X_1X_2}(11) = P_{X_1X_2}(00) = 5/16$ and $P_{X_1X_2}(01) = P_{X_1X_2}(10) = 3/16$. Therefore, the symmetry condition (13) is satisfied, which is a sufficient condition for zero secrecy leakage for two enrollments with the fuzzy commitment scheme. Also note that for this source-channel model, the secret-key capacity for each key is approximately $R_s = 0.0456$ bits/source-bit.

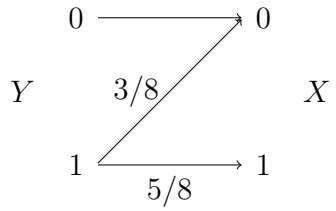


Figure 4: Z-channel with $z = 3/8$.

4 Conclusion

We have studied security of the helper data scheme in case of multiple enrollments. We proved that codes exist for any PUF source that is time and permutation invariant, such that the key remains secure when enrollment is repeated for a second time. Furthermore, we have shown that the fuzzy commitment scheme remains secure for any number of repeated enrollments when the PUF source meets a symmetry condition. The temperature-dependent model for SRAM-PUF meets the symmetry condition. We argued that many source-channel models exist that meet the symmetry condition and have shown examples.

Appendix A. Proof of Converse for Theorem 1

We show that for any number of enrollments, the secret-key rate cannot exceed $I(X_1; Z)$ for each secret key generated for the two-enrollment case. First, we have

$$\begin{aligned} H(S_t|\boldsymbol{Z}W_t) &= H(S_t|\boldsymbol{Z}W_t\widehat{S}_t) \\ &\leq H(S_t|\widehat{S}_t) \\ &\leq 1 + P_e \log_2 |\mathcal{S}_t| \leq 1 + \delta n \end{aligned} \tag{22}$$

where $P_e = \Pr(\widehat{S}_t \neq S_t) \leq \delta$ with $\delta > 0$. Then, the entropy of the key is

$$\begin{aligned} H(S_t) &= I(S_t; \boldsymbol{Z}W_t) + H(S_t|\boldsymbol{Z}W_t) \\ &\leq I(S_t; W_t) + I(S_t; \boldsymbol{Z}|W_t) + 1 + n\delta \\ &\leq H(\boldsymbol{Z}) - H(\boldsymbol{Z}|W_t S_t \boldsymbol{X}_t) + 1 + 2n\delta \\ &= nI(X_t; Z) + 1 + 2n\delta. \end{aligned} \tag{23}$$

This results in

$$R_t - \delta \leq \frac{1}{n} H(S_t) + \delta \leq I(X_t; Z) + \frac{1}{n} + 2\delta. \tag{24}$$

Now with $n \to \infty$ and $\delta \downarrow 0$ we obtain the proof of converse.

References

[1] R. Ahlswede and I. Csiszár, "Common randomness in information theory and cryptography - Part I: Secret sharing," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1121–1132, July 1993.

[2] U. M. Maurer, "Secret key agreement by public discussion from common information," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 2733–742, May 1993.

[3] A. Juels and M. Wattenberg, "A fuzzy commitment scheme," in *ACM Conf. Comp. Commun. Security*, New York, NY, Nov. 1999, pp. 28–36.

[4] L. Kusters, T. Ignatenko, and F. M. J. Willems, "Zero-leakage multiple key-binding scenarios for SRAM-PUF systems based on the XOR-method," in *6th Joint WIC/IEEE Symp. on Inf. Theory and Signal Proc. in the Benelux, May 19-20, 2016, Louvain, Belgium*, 2016, pp. 120–127.

[5] L. Kusters, T. Ignatenko, F. M. J. Willems, R. Maes, E. van der Sluis, and G. Selimis, "Security of helper data schemes for SRAM-PUF in multiple enrollment scenarios," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 1803–1807.

[6] O. Günlü, O. İşcan, and G. Kramer, "Reliable secret key generation from physical unclonable functions under varying environmental conditions," in *IEEE Int. Workshop Inf. Forensics Security*, Rome, Italy, Nov. 2015, pp. 1–6.

[7] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. John Wiley & Sons, 2006.

[8] R. Maes, "An Accurate Probabilistic Reliability Model for Silicon PUFs," in *Cryptogr. Hardw. Embed. Systems - CHES 2013 15th Int. Work. St. Barbar. CA, USA*, 2013, pp. 73–89.

[9] N. Chayat and S. Shamai, "Extension of an entropy property for binary input memoryless symmetric channels," *IEEE Trans. Inf. Theory*, vol. 35, no. 5, pp. 1077–1079, Sep. 1989.

[10] O. Günlü and G. Kramer, "Privacy, secrecy, and storage with multiple noisy measurements of identifiers," *IEEE Trans. Inf. Forensics and Security*, 2018, to appear.

Quantum Key Recycling with noise

Daan Leermakers and Boris Škorić
d.leermakers.1@tue.nl, b.skoric@tue.nl

Quantum cryptography uses the properties of quantum physics to achieve security feats that are impossible with classical communication. Best known is Quantum Key Distribution (QKD) famously described in the BB84 scheme. QKD establishes a random secret key known only to Alice and Bob, and uses the no-cloning theorem for unknown quantum states to detect any manipulation of the quantum states. Even before the invention of QKD, the idea of Quantum Key Recycling (QKR) was proposed. Let Alice and Bob encrypt classical data as quantum states, using a classical key to determine the basis in which the data is encoded. If they do not detect any manipulation of the quantum states, an attacker Eve has learned almost nothing about the encryption key, and hence it is safe for Alice and Bob to re-use the key. The idea of QKR, although proposed in 1982, was not proven to be secure. Only in 2005 Damgård, Pedersen and Salvail introduced a QKR scheme for which they could prove security. It does however need a quantum computer to the handle more involved quantum operations on highly dimensional quantum states used in the scheme. In 2017 attention returned to QKR based on prepare-and measure protocols acting on individual qubits, i.e. not requiring a quantum computer. Fehr and Salvail proved security in the almost noiseless case, and Škorić and de Vries introduced an encoding based on 8 states well suited for QKR.
We prove the security for QKR with 4-, 6- or 8-state encodings in the case of noisy channels. We take an 'engineering' point of view: we do not aim for complete key re-use, but rather for a high ratio of message length versus expended key bits. Our proof is based on a computation of the statistical distance between the real quantum state of the system and a state in which the secrets are completely secure. (Apart from the messages, also the keys have to stay secret. The worst case attack is a known-plaintext attack on the keys followed by an attack on the last message.) We can interpret this as being ideally secure except with negligible probability. This means that except for the small probability, an adversary starts the 2nd round with zero information about the keys. This argument can be repeated to prove the security of the keys for an arbitrary number (N) of rounds before refreshing the key material. When the key is completely unknown to an attacker, no information can be gained about the message in the case of 8-state encoding. The 4- and 6-state encodings need some extra privacy amplification.
We define the communication rate as the difference between the length of the message communicated and the key material spent, per round and per qubit. Let n be the number of qubits sent. Let β be the bit error rate on the quantum channel. Given N rounds without rejections, the asymptotic ($n \gg 1$) rate for 8-state encoding is given by

$$\frac{|\text{total msg}| - |\text{total key}|}{Nn} = 1 - h(\beta) - 2\log\left[\sqrt{(1-\beta)(1-\tfrac{3}{2}\beta)} + \sqrt{\tfrac{1}{2}\beta(1-\beta)} + \beta\sqrt{2}\right]. \qquad (1)$$

The asymptotic rate is positive up to $\beta \approx 0.09$. This result is better than expected given the known min-entropy result [1]. The full version of the paper can be found online [2].

References

[1] D. Leermakers and B. Škorić. Optimal attacks on qubit-based Quantum Key Recycling. *Quantum Information Processing*, 2018.

[2] D. Leermakers and B. Škorić. Security proof for quantum key recycling with noise. *https://eprint.iacr.org/2018/264.pdf*, 2018.

Round Robin Differential Phase Shift QKD security proof

Daan Leermakers and Boris Škorić
TU Eindhoven
d.leermakers.1@tue.nl, b.skoric@tue.nl

Introducton - Quantum cryptography uses the laws of quantum physics to perform cryptographic tasks. A measurement of a quantum system typically destroys information; it is impossible to clone unknown quantum states. These properties make it possible to achieve security features that are impossible in classical cryptography. The best known type of quantum cryptography is Quantum Key Distribution (QKD) introduced by Bennett and Brassard in 1984. In QKD two parties, Alice and Bob, create a shared secret key completely unknown to an attacker Eve. In BB84 this is achieved by encoding classical information in the polarization of photons in one of two non-orthogonal bases. The fact that the states are non-orthogonal makes it impossible for Eve to perfectly distinguish between them without any side information. QKD is by no means limited to two-dimensional Hilbert spaces or photon polarization. Since the invention of BB84 many QKD have been proposed, Round-Robin Differential Phase Shift (RRDPS) QKD is a particular interesting one. Introduced by Sasaki, Yamamoto and Koashi in 2014 [1], RRDPS is a QKD scheme based on d-dimensional qudits. It encodes the information in the phase difference between photons in an arbitrarily long pulse train. It has the advantage of having positive rate even when using very noisy quantum channels. In this paper we prove the security of the RRDPS protocol against general attacks.

The protocol

1. Alice generates a random bitstring $a \in \{0,1\}^d$. She encodes this random information in the phases between the pulses in a pulse train by preparing the single-photon state
$$|\mu_a\rangle \stackrel{\text{def}}{=} \frac{1}{\sqrt{d}} \sum_{t=0}^{d-1} (-1)^{a_t} |t\rangle. \tag{1}$$
She sends the state to Bob.
2. Bob measures the phase difference between two random arrival times of the photon using a variable-delay interferometer. He chooses a random integer $r \in \{1, \ldots, d-1\}$ and performs the POVM measurement $\mathcal{M}^{(r)}$
$$M_{ks}^{(r)} = \frac{1}{2} |\Psi_{ks}^{(r)}\rangle\langle\Psi_{ks}^{(r)}| \qquad |\Psi_{ks}^{(r)}\rangle = \frac{|k\rangle + (-1)^s |k+r\rangle}{\sqrt{2}}. \tag{2}$$
The result of the measurement $\mathcal{M}^{(r)}$ on $|\mu_a\rangle$ is an random integer $k \in \{0, \ldots, d-1\}$ and a bit $s = a_k \oplus a_{k+r}$.
3. Bob announces k and r over a public but authenticated channel. Alice computes $s = a_k \oplus a_{k+r}$. Alice and Bob now have a shared secret bit s.

This procedure is repeated multiple times. Finally, Alice and Bob carry out the standard procedures of information reconciliation and privacy amplification.
To determine the bit error rate, Alice and Bob have to compare a randomly selected fraction of their secret bits. If this comparison is not performed, Alice and Bob have to assume that Eve learns as much as when causing bit error rate $\beta = \frac{1}{2}$. Due to the noise resistance of the protocol, security can be proven up to arbitrarily high noise levels.

Prior work - In [1], an upper bound for asymptotic key length was computed. Eve's information about the shared key is bounded by
$$I_{\text{AE}} \leq h(\frac{1}{d-1}) \tag{3}$$
(Eq. 5 in [1] with photon number set to 1). The security analysis is based on an entropic inequality for non-commuting measurements. There are two issues with this analysis. First, the proof is not written out in detail. Second, it is not known how tight the bound is.

Ref. [2] follows [1] and does a more accurate computation of phase error rate, tightening the $1/(d-1)$ in (3) to $1/d$. In [3] Sasaki and Koashi add β-dependence to their analysis and claim a bound

$$I_{\mathrm{AE}} \leq h(\frac{2\beta}{d-2}) \qquad \text{for } \beta \leq \frac{1}{2} \cdot \frac{d-2}{d-1} \tag{4}$$

and $I_{\mathrm{AE}} \leq h(\frac{1}{d-1})$ for $\beta \in [\frac{1}{2} \cdot \frac{d-2}{d-1}, \frac{1}{2}]$.

In this paper we improve upon these bounds. We show that less privacy amplification is needed and that saturation of the amount of privacy amplification occurs at lower noise levels than suggested in prior work. In addition we give a clear description of the restrictions the protocol forces onto Eve's state and we are able to prove the security using the well know instrument of statistical distance and (conditional) Von Neumann entropy.

Proof structure - We first analyse an attack in which Eve couples an ancilla to each EPR pair individually with no restrictions other than the noise parameter β. Post-selection [4] can be used to show this is sufficient to prove security again general attacks as long as the number of exchanged qudits n is chosen such that $2d^4 \log n \ll n$. We consider the RRDPS scheme with channel monitoring. Alice and Bob test the bit error rate *for each combination* (a, k) *separately*, demanding that for each (a, k) the observed bit error rate does not exceed β. To facilitate this, we will assume that n is chosen sufficiently large to ensure $d2^d \log n \ll n$. If channel monitoring is not performed $\beta = \frac{1}{2}$ has to be assumed.

We show that the RRDPS protocol is equivalent to a protocol that contains an additional randomisation step by Alice and Bob. Here equivalence means all variables and measurement outcomes have the same probability distributions as the original protocol. The randomisation consists of phase flips and a permutation of the basis states. We construct an EPR variant of RRDPS-with-randomisation; it is equivalent to RRDPS if Alice creates the EPR pair and immediately does her measurement. The effect of the randomisation is that Alice and Bob's entangled state after Eve's attack on the EPR pair is symmetrised. We then impose the noise constraint on the state shared by Alice and Bob and describe Eve's state by constructing the purification of their state. This corresponds to the worst case assumption that all information that leaks from Alice and Bob's shared state is in Eve's hands. Eve's resulting state has only two degrees of freedom.

We compute the statistical distance between the real protocol and an idealized version of the protocol in which Eve is completely decoupled from the final key. This results in an expression for the length of the final shared key which also holds in the non-asymptotic regime. For very large n, the Von Neumann entropy is a sharper measure for the security i.e. it results in a smaller amount of privacy amplification. For both expressions of security, the two degrees of freedom left in Eve's ancilla state are optimized.

To gain additional insight in Eve's optimal attack, we use the POVM formalism to describe Eve's optimal measurement on the ancilla coupled to a single qudit in terms of min-entropy and accessible information. In this restricted class of attacks, saturation already occurs at $\beta = \frac{1}{4} \cdot \frac{d-2}{d-1}$ i.e. at half the value of [3].

Main results - Figure 1 shows the required amount of privacy amplification in the case of saturated noise levels. In that case there is no need for Alice and Bob to sacrifice a fraction of the qubits for channel monitoring purposes. We show the two main results of the paper as well as the bound given in [1]. Figure 2 shows the same bounds as a function of the noise parameter β for fixed d=16. Here we compare with the known asymptotic result in [3].

Our results show improvements over previous bounds in the asymptotic regime as well as providing bounds that hold in the non-asymptotic case. Remarkably even the non-asymptotic bounds are sharper under some conditions than the previous asymptotic bounds.

The full version of this paper can be found online [5].

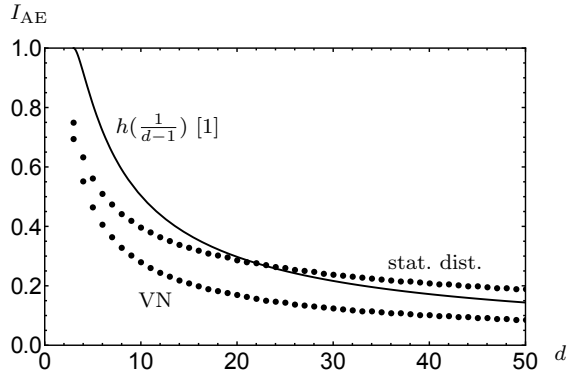


Fig. 1 *Saturated leakage as a function of d. Comparison of [1] and our results for the statistical distance and the Von Neumann entropy.*

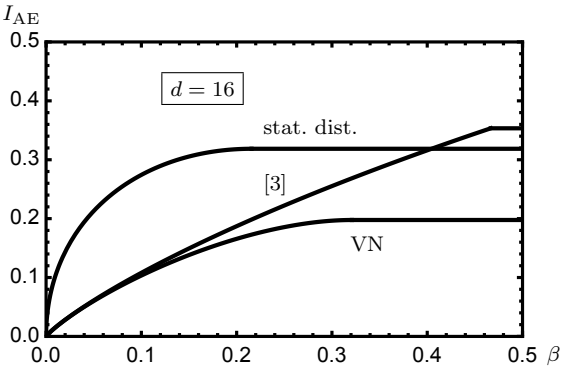


Fig. 2 *Leakage as a function of β, for $d = 16$. Comparison of our results for the statistical distance and the Von Neumann entropy versus equation 4 [3].*

References

1. T. Sasaki, Y. Yamamoto, and M. Koashi. Practical quantum key distribution protocol without monitoring signal disturbance. *Nature*, 509:475–478, May 2014.
2. Z. Zhang, X. Yuan, Z. Cao, and X. Ma. Round-robin differential-phase-shift quantum key distribution. `http://arxiv.org/abs/1505.02481v1`, 2015.
3. T. Sasaki and M. Koashi. A security proof of the round-robin differential phase shift quantum key distribution protocol based on the signal disturbance. `https://arxiv.org/abs/1701.08509`, 2017.
4. Matthias Christandl, Robert König, and Renato Renner. Postselection technique for quantum channels with applications to quantum cryptography. *Phys. Rev. Lett.*, 102:020504, Jan 2009.
5. D. Leermakers and B. Škorić. Security proof for Round Robin Differential Phase Shift QKD. `https://eprint.iacr.org/2017/830.pdf`, 2018.

Improved BER Performance of Hard-decision Staircase Code via Geometric Shaping

Yi Lei[1], Bin Chen[1,2], and Alex Alvarado[1]
[1]Signal Processing Systems Group, [2]Electro-Optical Communications Group
Department of Electrical Engineering, Eindhoven University of Technology,
Eindhoven 5600 MB, The Netherland
{y.lei,b.c.chen,a.alvarado}@tue.nl

Abstract

Staircase codes (SCCs) with hard decision (HD) decoding have attracted much attention in the optical communication community due to their outstanding performance and low decoding complexity [1]. Recently, an implementation agreement has been reached for using an SCC as an outer code in the baseline draft of 400G ZR [2]. To achieve high spectral efficiencies, forward error correction is typically combined with high order modulation formats. Additional gains can be obtained if regular quadrature amplitude modulation (QAM) constellations are replaced by constellations with geometric shaping (GS). In this paper, we consider GS with HD-SCC and analyze the post-SCC bit error rate (BER) performance of constellations with 64 points. Bose-Chaudhuri-Hocquenghem (BCH) codes are used as SCC component codes. The location of the constellation points is optimized by maximizing the achievable information rate. The simulation results show that the shaped constellations yield around 0.24 dB gains at a BER of 10^{-6} when compared to regular 64QAM for different coding rates.

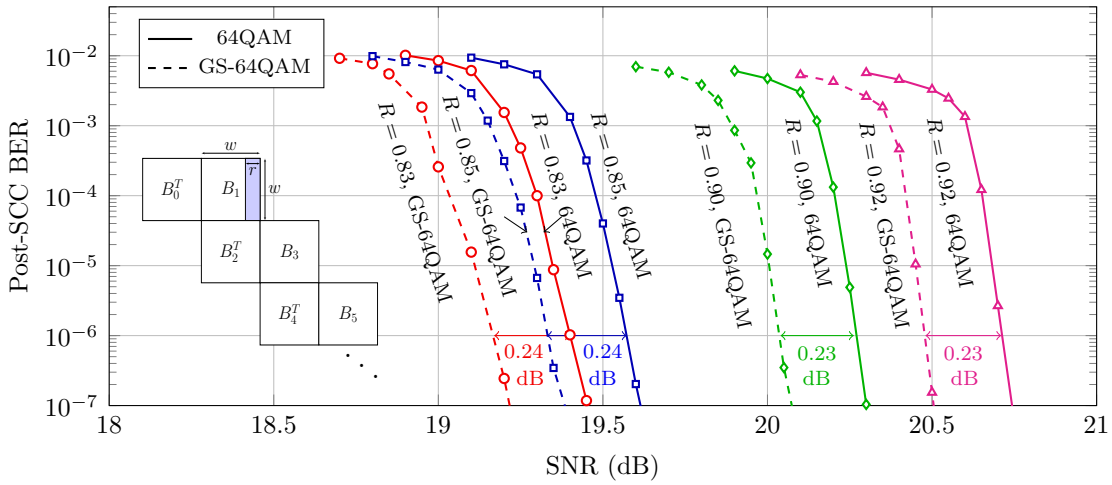


Figure 1: BER vs. SNR for an AWGN channel. The inset shows the "staircase" structure of the SCC. Each SCC block size is $w \times w$, including $w \times (w-r)$ information bits and $w \times r$ parity bits. Each row of $[B_{i-1}^T B_i]$ is a valid codeword in the component BCH code. For coding rates $R = 0.83, 0.85, 0.90$ and 0.92, the corresponding SCC block sizes are 114×114, 126×126, 192×192 and 252×252.

References

[1] Benjamin P. Smith, Arash Farhood, Andrew Hunt, Frank R. Kschischang, and John Lodge, "Staircase Codes: FEC for 100 Gb/s OTN," J. Lightw. Technol., pp. 110-117, Jan. 2012.

[2] Optical Internetworking Forum, "Implementation Agreement 400ZR," OIF 400G ZR, Jan. 2018.

The Behavior of Principal Component Analysis and Linear Discriminant Analysis (PCA-LDA) for Face Recognition

Nova Hadi Lestriandoko[1,2] Luuk Spreeuwers[1] Raymond Veldhuis[1]
[1]University of Twente
Faculty of EEMCS, DMB group
Netherlands
[2]Indonesian Institute of Sciences
Research Center for Informatics
Indonesia
[1](n.h.lestriandoko; l.j.spreeuwers; r.n.j.veldhuis)@utwente.nl
[2]nova002@lipi.go.id

Abstract

This paper presents the analysis of PCA-LDA behavior for face recognition using Singular Value Decomposition (SVD). The experimental results is shown to analyze face recognition performance, i.e. the impact of number of subjects, images per subject, training set size, and trade-off between the number of subjects and the number of images per subject on recognition performance, in relation with the number of PCA-LDA coefficients. The comparison of three classifiers, i.e. Euclidean Distance, Cosine Similarity, and Likelihood Ratio, are presented to obtain knowledge about their characteristics. All experimental evaluations are in the verification context. Based on the experimental results, the larger number of subjects and images per subject produced the better recognition performance. Regarding the number of subjects and images per subject trade-off, its indicated both of them influence the recognition performance. Otherwise, the image size also affect to recognition performance. PCA-LDA can perform low resolution image well up to 15x15 pixels and breaks down afterward. Regarding the p and ℓ coefficients, PCA-LDA has different behavior for each classifier.

1 Introduction

Face recognition based on eigenfaces or Principal Component Analysis (PCA) was introduced by M.Turk and A.Pentland [1]. This method reduced the dimensionality by transforming the features from a higher dimensionality space to a lower dimensionality space. The PCA projects face images onto a feature space spanned by the eigenfaces that are the eigenvectors of the covariance matrix of the vector space of face images. The recognition is performed by measuring similarity using classification techniques. Still in the linearly projection area, Fisherfaces or Linear Discriminant Analysis (LDA), introduced by Belhumeur et al[2], maximizes the ratio of the within class and the between class to obtain the best separation of the classes. The dimensionality reduction is done by choosing the most significant coefficients: p largest PCA eigenvalues and l smallest LDA eigenvalues from remaining p.

Commonly, there are many LDA improvements implemented in Biometrics applications. Veldhuis et al.[3] demonstrated the feasibility of hand-geometry recognition based on contour parameters. The integrating LDA with likelihood ratio as classifier

was presented in Veldhuis[4] and Spreeuwers et al[5]. Moreover, Sharma et al.[6] proposed a two-stage linear discriminant analysis technique that regularize the between-class scatter and within-class scatter matrices in parallel to produce two orientation matrices, which is concatenated afterward. Still in the context of LDA improvement, Ioeffe [7] presented the probabilistic LDA by modelling both within class and between class variations to solve recognition problems on classes that unseen before. On the other hand, in the context of LDA performance analysis, Zanetti et al.[8] presented the reports on the impact of the number of individuals, the number of images per individual, and trade-off between them to face recognition. However, there are no explanation about PCA-LDA behavior in relation with the number of coefficients.

This paper presents the behavior of PCA-LDA on three classifiers and analyzes the effect of the number of subjects, the number of images per subject, images size, and trade-off between the number of subjects and the number of images per subject on recognition performance. This paper is organized as follows: Section 2 presents the PCA for face recognition. Section 3 continues to LDA and its implementation. Section 4 deals with Similarity Score. The experiments and results are showed in the section 5 and 6, followed by the conclusion at the end of paper.

2 Principle Component Analysis for Face Recognition

PCA is a statistical approach, introduced by M.Turk *et al* [1], used to extract the most relevant features to describe faces. In PCA, every image in the training set is represented as a linear combination of weighted eigenvectors called eigenfaces. PCA can be written as:

$$\mathbf{M_t} \approx \mu_\mathbf{t} + \sum_{i=1}^{k} \mathbf{u_i}\omega_i \tag{1}$$

where k is the number of eigenfaces (eigenvectors) and $k \leq d$. Then, the weight ω_i can be computed easily because of orthonormality as:

$$\omega_j = \mathbf{u_j}^T(\mathbf{M_t} - \mu_\mathbf{t}) \tag{2}$$

where ω_j are weighted features, $\mathbf{u_j}$ are the eigenvectors, $\mathbf{M_t}$ are facial images with dimensionality d, and $\mu_\mathbf{t}$ is the average of the training set.

These eigenvectors are obtained from the covariance matrix of a training set. The weights are obtained after selecting a set of most relevant Eigenfaces. For the verification recognition, the score is obtained by projecting a test image onto the subspace spanned by the eigenfaces and then classification is done by calculating the similarity score.

2.1 Training

Below are the steps to train PCA as feature extraction for face recognition:

1. Prepare the training faces.
   Obtain face images, preprocess them to get centered face in the same size with dimension d.

2. Prepare dataset.
   For every face image in the database, transform into a vector $\mathbf{m_i}$ with size $d \times 1$ and place into a training set $\mathbf{M_t}$.

$$\mathbf{M_t} = \{\mathbf{m_1}, \mathbf{m_2}, \mathbf{m_3}, ..., \mathbf{m_n}\} \tag{3}$$

where n is the number of training samples. So, $\mathbf{M_t}$ is an $d \times n$ matrix.

3. Compute the average of face vector.
The average of face vector $\mu_\mathbf{t}$ can be calculated by using the following formula:

$$\mu_\mathbf{t} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m_i} \tag{4}$$

4. Centering the data (subtract the average face vector).
To get centered data, the face vector is subtracted by the average of face vector. Then, the result is stored in $\mathbf{Z}$.

$$\mathbf{Z_i} = \mathbf{m_i} - \mu_\mathbf{t} \tag{5}$$

and for matrix $\mathbf{A}$:

$$\mathbf{A} = \{\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{Z_3}, ..., \mathbf{Z_n}\} \tag{6}$$

Now, the size of $\mathbf{A}$ is $d \times n$.

5. Calculate the covariance matrix.
The covariance matrix can be obtained by following formula:

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{Z_i}\mathbf{Z_i}^T = \frac{1}{n-1}\mathbf{A}\mathbf{A}^T \tag{7}$$

Now the size of matrix $\mathbf{C}$ is $d \times d$.

6. Calculate the Eigenvectors and Eigenvalues.
The matrix $\mathbf{C}$ has size $d \times d$, so it will have d eigenvalues. For this case, the computationally intensive is very hard because of a large dimensional matrix.

PCA reduces the dimensionality by calculating eigenvectors of matrix $\mathbf{A}^T\mathbf{A}$. Both of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ have the same eigenvalues λ. If $\mathbf{u_i}$ is eigenvector of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{v_i}$ is eigenvector of $\mathbf{A}^T\mathbf{A}$, then the relation of $\mathbf{u_i}$ and $\mathbf{v_i}$ is described in the following equations:

$$\mathbf{A}^T\mathbf{A}\mathbf{v_i} = \lambda_i\mathbf{v_i} \tag{8}$$

Multiplying both sides by $\mathbf{A}$,

$$\mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{v_i} = \lambda_i\mathbf{A}\mathbf{v_i} \tag{9}$$

From this equation, $\mathbf{A}\mathbf{v_i}$ are the eigenvectors of $\mathbf{C} = \mathbf{A}\mathbf{A}^T$ and both of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ have same eigenvalues λ_i, thus:

$$\mathbf{u_i} = \mathbf{A}\mathbf{v_i} \tag{10}$$

We can use Singular Value Decomposition (SVD), that is a robust approach, to calculate eigenvalues and eigenvectors. SVD is a decomposition of a real or complex matrix that factorize a matrix into three matrices $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$. The columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal and the matrix $\mathbf{S}$ is a diagonal matrix with positive real entries. Both columns of $\mathbf{U}$ and $\mathbf{V}$ form an orthogonal set. The matrices $\mathbf{U}$ and $\mathbf{V}$ are also called left singular vectors and right singular vectors. The SVD theorem states:

$$\mathbf{A}_{nxd} = \mathbf{U}_{nxn}\mathbf{S}_{nxd}\mathbf{V}^T_{dxd} \tag{11}$$

where

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}_{nxn} \tag{12}$$

$$\mathbf{V}^T\mathbf{V} = \mathbf{I}_{dxd} \tag{13}$$

Calculating the SVD consists of finding the eigenvalues and eigenvectors of $\mathbf{AA}^T$ and $\mathbf{A}^T\mathbf{A}$:

(a) The eigenvectors of $\mathbf{A}^T\mathbf{A}$ make up the columns of $\mathbf{V}$,

(b) The eigenvectors of $\mathbf{AA}^T$ make up the columns of $\mathbf{U}$.

(c) The singular values in $\mathbf{S}$ are square roots of eigenvalues from $\mathbf{AA}^T$ or $\mathbf{A}^T\mathbf{A}$.

The singular values of matrix $\mathbf{S}$ is a diagonal matrix and arranged in descending order. The singular values are always real numbers. If the matrix $\mathbf{A}$ is a real matrix, then $\mathbf{U}$ and $\mathbf{V}$ are also real.

7. Choose only K eigenvectors corresponding to the K largest eigenvalues and project into eigenspace.

2.2 Recognition Procedure

Face recognition can be done by projecting a new facial image onto eigenspace by following formula:

$$\omega_i = \mathbf{u_i}^T(\mathbf{M_{new}} - \mu_\mathbf{t}) \tag{14}$$

where $i = 1, 2, 3, .., K$ and u_i is the eigenvectors corresponding with K largest eigenvalues. The last step of PCA feature extraction is to form feature vector:

$$\mathbf{\Omega_t} = [\,\omega_1 \quad \omega_2 \quad \omega_3 \quad \dots \quad \omega_K\,]^T \tag{15}$$

Finally, the recognition is done by calculating the similarity score between two feature vectors and comparing two thresholds.

3 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) based face recognition was introduced by Belhumeur *et al.* [2]. The LDA aim is to find out the best projection of original data matrix on a lower dimensional space by maximizing the ratio of the within class and the between class variances. The technique is to model the space and make it feasible using PCA, then transform the space using LDA in such away that the components are ordered with respective discriminative properties. The feature reduction is done by discarding the least discriminative components. The figure 1 shows how LDA works by projecting the dataset into two rotates axes. Projection to the lower right axis achieves the maximum separation between the categories and projection to the lower left axis yields the worst separation. There are some publications [7][9][10][11] that present the detail tutorial of LDA and its implementation for face recognition. They also described some weaknesses of LDA and introduced their solutions.

PCA-LDA can be implemented using SVD decomposition[3][5][12]. The sub section below will discuss how PCA and LDA produce a transformation matrix to obtain a feature vector.

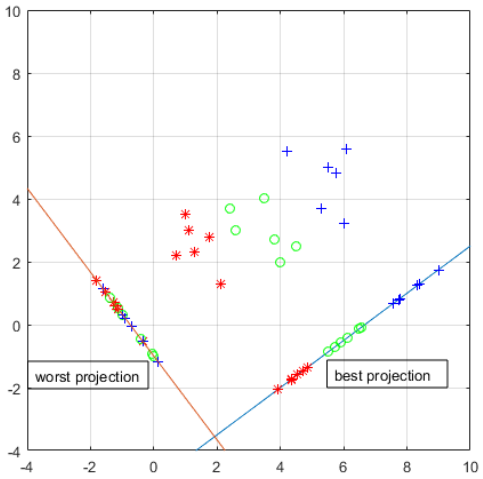


Figure 1: LDA projection

3.1 PCA Transformation

Suppose we have a training set of data consist of n face images. Then, the n sample vectors $\mathbf{m_i}$ of the face images with dimension d are ordered in an $n \times d$ matrix $\mathbf{M}$. If the zero mean of matrix $\mathbf{M}$ is $\mathbf{Z}$ and the mean of $\mathbf{M}$ is $\mu_\mathbf{t}$, then the covariance matrix of $\mathbf{Z}$ is defined by:

$$\mathbf{C_t} = \frac{1}{n-1}\mathbf{Z}\mathbf{Z}^T \tag{16}$$

The SVD decomposition of $\mathbf{Z}$ is:

$$\mathbf{Z} = \mathbf{U_t}\mathbf{S_t}\mathbf{V_t}^T \tag{17}$$

Where $\mathbf{U_t}$ and $\mathbf{V_t}$ are the left and right singular vectors (i.e. the eigenvectors of $\mathbf{Z}\mathbf{Z}^T$ and $\mathbf{Z}^T\mathbf{Z}$ respectively) and unitary matrices (i.e. $\mathbf{U_t}\mathbf{U_t}^T = \mathbf{I}$ and $\mathbf{V_t}\mathbf{V_t}^T = \mathbf{I}$). $\mathbf{S_t}$ is a diagonal matrix with the singular values of $\mathbf{Z}$, which are the square roots of the eigenvalues of $\mathbf{Z}\mathbf{Z}^T$. Then, the equation of $\mathbf{C_t}$ becomes:

$$\mathbf{C_t} = \frac{1}{n-1}\mathbf{Z}\mathbf{Z}^T = \mathbf{U_t}\frac{\mathbf{S_t}\mathbf{S_t}}{n-1}\mathbf{U_t}^T = \mathbf{U_t}\mathbf{\Sigma_t}\mathbf{U_t}^T \tag{18}$$

So, the transformation that whitens the total distribution is:

$$\mathbf{T_1} = \sqrt{n-1}\mathbf{S_t}^{-1}\mathbf{U_t}^T \tag{19}$$

Because:

$$\mathbf{T_1}\mathbf{C_t}\mathbf{T_1}^T = \mathbf{I} \tag{20}$$

The next step is to transform all sample vectors $\mathbf{m_i}$ by $\mathbf{T_1}$.

$$\mathbf{m_i'} = \mathbf{T_1}(\mathbf{m_i} - \mu_\mathbf{t}) \tag{21}$$

3.2 LDA Transformation

We assume that the within distribution of all subjects is normal with the same within class covariance $\mathbf{C_w}$, but different means and that the total distribution of all faces is normally distributed with total covariance $\mathbf{C_t}$. The within class covariance $\mathbf{C_w}$ are then calculated from the transformed data. If there are various classes c in the sample vectors and each class has n_c samples, so the mean of each class can be obtained by summing over the sample vectors of the specic class and dividing by the number of samples of the specic class:

$$\mu_{\mathbf{ci}} = \frac{1}{n_c} \sum_{i:m_i \in c} \mathbf{m_i'} \tag{22}$$

So, the zero mean of vector $\mathbf{m_i'}$ is:

$$\mathbf{z_{ci}} = \mathbf{m_i'} - \mu_{\mathbf{ci}} \tag{23}$$

The within class covariance $\mathbf{C_w}$ can be estimated by ordering the class zero mean vector $\mathbf{z_{ci}}$ into an $n \times d$ matrix $\mathbf{Z_c}$:

$$\mathbf{C_w} = \frac{1}{n-1} \mathbf{Z_c} {\mathbf{Z_c}}^T \tag{24}$$

Using SVD decomposition, we can obtain:

$$\mathbf{C_w} = \frac{1}{n-1} \mathbf{Z_c} {\mathbf{Z_c}}^T = \mathbf{U_w} \frac{\mathbf{S_w} \mathbf{S_w}}{n-1} {\mathbf{U_w}}^T = \mathbf{U_w} \mathbf{\Sigma_w} {\mathbf{U_w}}^T \tag{25}$$

And the transformation that decorrelates the within distribution is:

$$\mathbf{T_2} = {\mathbf{U_w}}^T \tag{26}$$

Finally, the total transformation $\mathbf{T}$ is the product of the two transformations:

$$\mathbf{T} = \mathbf{T_2} \mathbf{T_1} = {\mathbf{U_w}}^T \sqrt{n-1} {\mathbf{S_t}}^{-1} {\mathbf{U_t}}^T \tag{27}$$

The best discrimination in LDA is obtained by projecting the vectors on the subspace with the ℓ smallest eigenvalues. If the dimensionality of first transformation ($\mathbf{T_1}$) is reduced to p, then the dimensionality of second transformation ($\mathbf{T_2}$) is reduced to ℓ. This means only the smallest ℓ eigenvalues of the p remaining eigenvalues and corresponding eigenvectors are used resulting the final transformation ($\mathbf{T}$). So, the optimum performance is obtained by seeking the ℓ smallest eigenvalues from the best p largest eigenvalues or we can write:

$$\mathbf{T_{p\ell}} = \mathbf{T_{2\ell}} \mathbf{T_{1p}} = {\mathbf{U_{w\ell}}}^T \sqrt{n-1} {\mathbf{S_{tp}}}^{-1} {\mathbf{U_{tp}}}^T \tag{28}$$

Where $\mathbf{T_{p\ell}}$ is a $d \times \ell$ transformation matrix. Thus, the features can be extracted using this transformation matrix.

4 Similarity Scores

There are so many methods to calculate similarity scores. This paper presents the use of Euclidean Distance, Cosine Similarity, and Likelihood Ratio as classifiers.

4.1 Euclidean Distance

In mathematics, Euclidean Distance is a distance between two points in the straight line. The calculation refers to the old literature as *Pythagorean* metric. So, the distance between two vectors $\mathbf{r}$ and $\mathbf{x}$ can be written as:

$$d(\mathbf{r}, \mathbf{x}) = \sqrt{\sum_i (r_i - x_i)^2} \tag{29}$$

4.2 Cosine Similarity

The cosine similarity measures the cosine of the angle between two vectors or points. The cosine similarity is defined as:

$$d(\mathbf{r}, \mathbf{x}) = \frac{\mathbf{r}^T \mathbf{x}}{||\mathbf{r}||||\mathbf{x}||} \tag{30}$$

4.3 Likelihood Ratio

The likelihood ratio can be regarded as a score and the decision if two facial images are of the same subject is taken by comparing this score to a threshold [3][5]. A simple expression for the log of the likelihood ratio can be calculated by first applying a transformation $\mathbf{T}$ that de-correlates and scales the total distribution such that it becomes white and simultaneously de-correlates the within distribution. This transformation is obtained using the singular values and vectors of $\mathbf{C_t}$ and $\mathbf{C_w}$. The derivation of the likelihood ratio, see [5], then becomes:

$$LR(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{\Lambda} \mathbf{x} + \mathbf{y}^T \mathbf{\Lambda} \mathbf{y} + (\mathbf{x} + \mathbf{y})^T \mathbf{\Gamma} (\mathbf{x} + \mathbf{y}) \tag{31}$$

Where:

$$\mathbf{\Lambda} = \mathbf{I} - {\mathbf{\Sigma_w}}^{-1} \tag{32}$$

$$\mathbf{\Gamma} = {\mathbf{\Sigma_w}}^{-1} [2{\mathbf{\Sigma_w}}^{-1} + {\mathbf{\Sigma_b}}^{-1}]^{-1} {\mathbf{\Sigma_w}}^{-1} \tag{33}$$

$$\mathbf{\Sigma_b} = \mathbf{I} - \mathbf{\Sigma_w} \tag{34}$$

All required parameters above are obtained from previous equation, i.e. $\mathbf{\Sigma_w}$ derivation in equation 25, see Linear Discriminant Analysis section.

5 Experiments

The experiments are conducted using FRGC v2 dataset. The FRGC consists of controlled and uncontrolled images. In these experiments, we used only controlled images. Firstly, the registration is applied to align the faces in the FRGC v2 dataset. It is an important preprocessing step for face recognition. The face properties, i.e. eyes, nose, mouth, and face, are detected automatically using Viola-Jones cascade detectors[13]. The registration refers to the eyes position and transform the face image using geometric transformation and linear interpolation. Moreover, the ellipse mask is applied to focus the recognition on face area without hair and ears. Then, the histogram equalization is performed to overcome the illumination problem[14]. The figure 2 shows the result of each face registration step.

The equal-error rate is used as verification rate for performance measurement. The four types of experiments were performed in different conditions and situations:

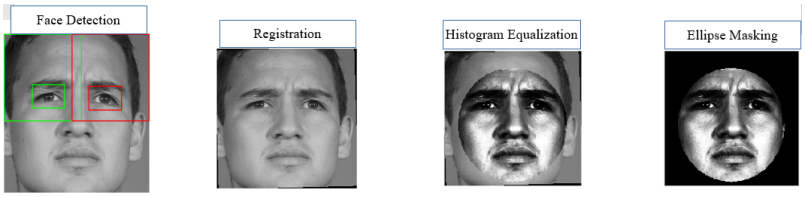


Figure 2: Face Registration

1. The Impact of Number of Subjects to Recognition Performance
   This experiment is used to know the impact of number of subjects to recognition performance, in relation with the behavior on three classifiers: Euclidean Distance, Cosine Similarity, and Likelihood Ratio. The dataset was separated into 3000 samples for training and 500 samples for testing with image size 50x50 pixels. First, we took 500 samples from 25 subjects to train the system. Then, the classifier parameters and the transformation matrix was calculated from training set and applied into testing set. So, the similarity scores were obtained from testing set. The next step was to extend the number of subjects on training set, i.e. 1000, 2000, and 3000 samples. The behavior of PCA-LDA was obtained by analyzing p and ℓ coefficients on three classifiers.

2. The Impact of Number of Images per Subject
   The training set was divided into 4 subsets to analyze the impact of number of images per subject. We took 25 subjects with various numbers of images per subject, i.e. 2, 5, 10, and 20 images per subject respectively. The recognition performance was obtained from testing set with 500 samples from 25 subjects. The likelihood ratio classifier was used in this experiment. Based on the results of experiment 1 on 500 training set and likelihood classifier, the p and ℓ coefficients is defined to 28 and 23.

3. The Impact of Number of Images per subject and Number of Subjects trade off
   The third experiment was conducted to find a trade off between number of images per subject and number of subjects. In this experiment, we used 400 training set from various subjects(i.e. 200, 80, 40, and 20) and various images number per subject(i.e. 2, 5, 10, 20). The p coefficient is defined to 200 and the observation is done along ℓ axis.

4. The Impact of Image Size
   An image contains information as much as its resolution or size. The aim of this experiment is to know the impact of image size, i.e. 200x200, 100x100, 50x50, 40x40, 30x30, 25x25, 20x20, 15x15, 12x12, and 10x10, to recognition performance. The 2000 training set from 100 subjects and the 500 testing set from 25 subjects were used to verify the performance. The PCA-LDA coefficients are defined as: p=100 and ℓ=25.

6 Results

6.1 The Impact of Number of Subjects to Recognition Performance

The results showed a valley, as an optimal area, along ℓ coefficient axis at $\ell \leq s - 1$ on Euclidean and Cosine classifier. In this case, s represents the number of subjects. The valley became smaller for bigger training set and disappeared for training set more

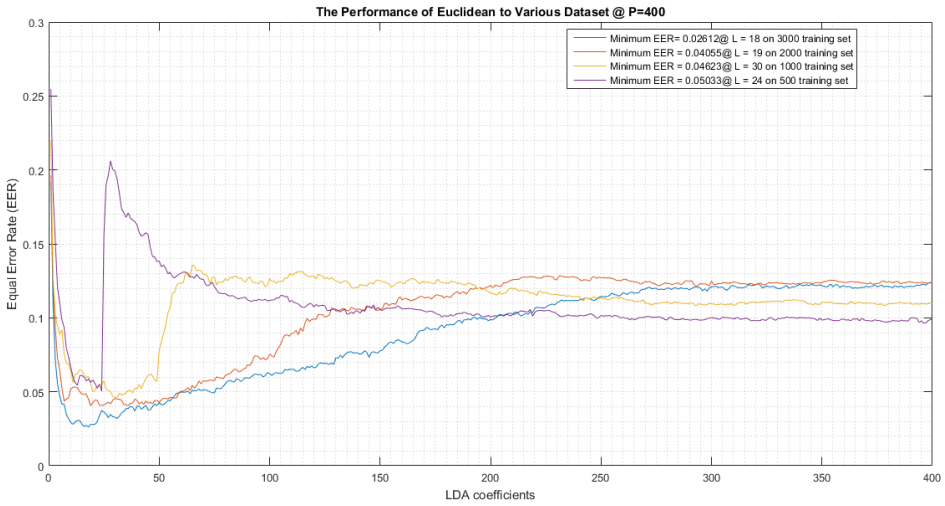


Figure 3: LDA coefficients using Euclidean Distance on various dataset at p=400

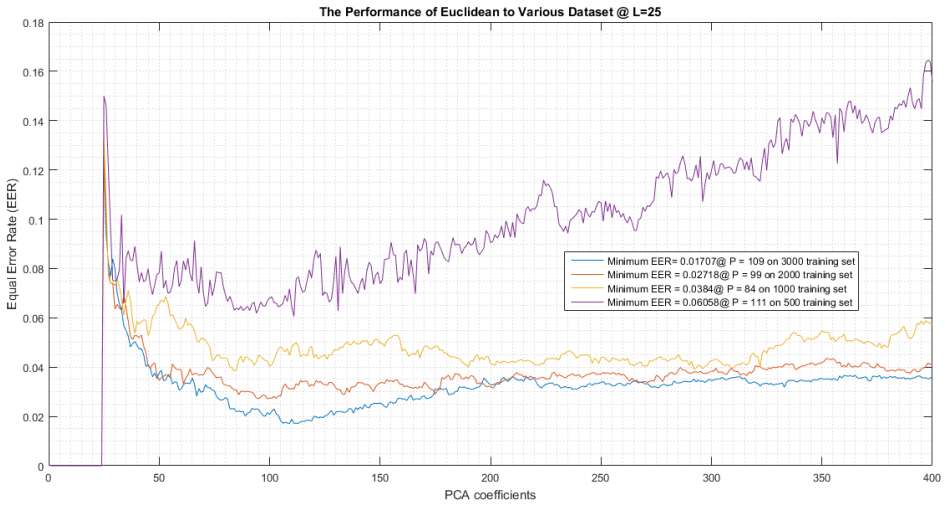


Figure 4: LDA coefficients using Euclidean Distance on various dataset at ℓ=25

than 2000. Similarly, PCA-LDA with likelihood ratio has optimal area at $\ell \leq s-1$. After that point, the performance line becomes flat or there are no contribution to recognition. It is happened because PCA-LDA maximizes the ratio of the between class and the within class which has the optimum dimension at $d \leq (p, s-1)$ [3]. The detail explanation for each classifier can be seen in the sub section below.

6.1.1 Euclidean Distance

Based on the performance observation on all parameters, we obtained the behavior of PCA-LDA on Euclidean Distance classifier. The results showed a valley with a bump at LDA coefficient $\ell = s-1$. figure 3 showed the bump position at ℓ coefficient on dataset with various number of subjects. The bump size is vice versa with the number of subjects. The bigger the number of subject, the smaller the bump size. But, the

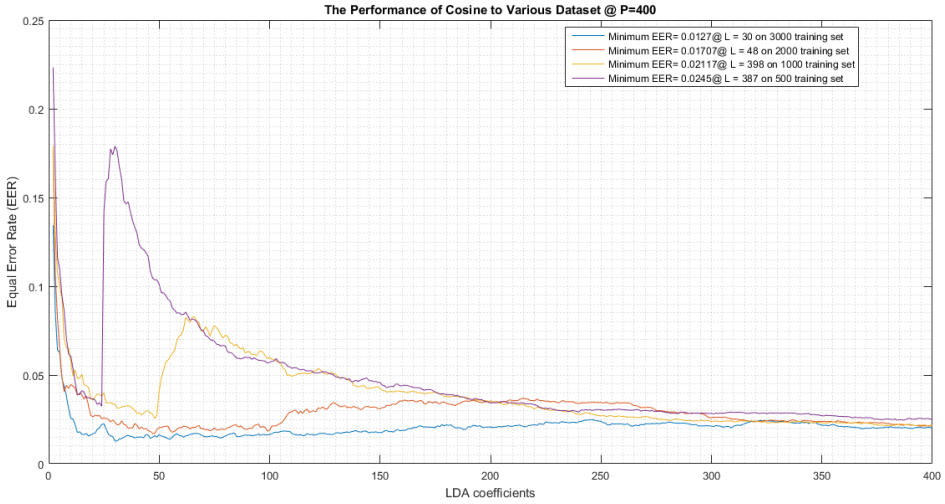


Figure 5: LDA coefficients using Cosine Similarity on various dataset at p=400

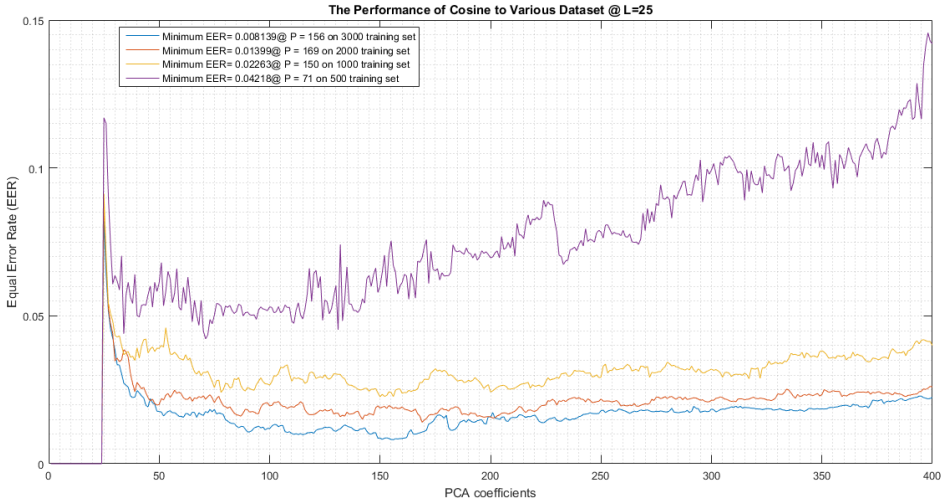


Figure 6: LDA coefficients using Cosine Similarity on various dataset at ℓ=25

bump disappeared on 3000 training set. It means that we have to provide minimum 2000 training set to get a representative training set. However, the optimal performance was always in the range of ℓ =10 to 50.

The next observation for p coefficient, we took ℓ=25 and observed the performance for all p coefficients. As shown in figure 4, the graphics showed that the trend of optimal performance was in the range p=80 to 120. The number of subjects also gave an impact to recognition performance. The 500 training set produced high fluctuation and the worst performance. Otherwise, the larger the training set, the better and the more stable the performance.

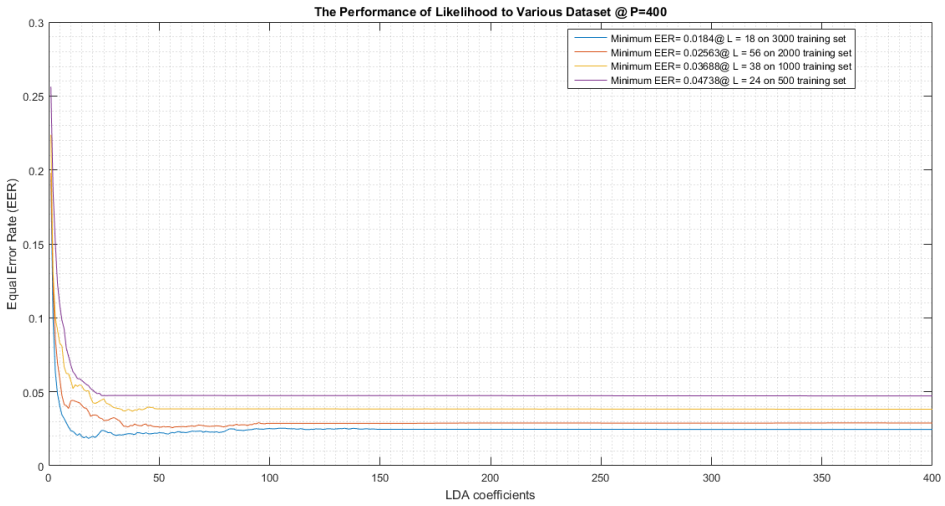


Figure 7: LDA coefficients using Likelihood Ratio on various dataset at p=400

6.1.2 Cosine Similarity

Similar with the Euclidean Distance, The LDA coefficient observation showed that Cosine has a bump at LDA coefficient $\ell = s - 1$ and the bump size becomes smaller on bigger dataset. Whereas, the ℓ observation at p=400 is shown in the figure 5. The optimal area was in the range ℓ=10 to 100. But, the trend on all dataset went down to the minimum EER after the bump. For example, the minimum EER on 500 dataset was reached at ℓ=387 and 1000 dataset at ℓ=398. It was possible for all dataset to have lower EER at higher ℓ coefficient, because the limitation of our observation was 400 coefficients.

Regarding the observation on p coefficient, as shown in the figure 6, the optimal performance for Cosine was reached at p=70 to 200. The graphics showed the performance of Cosine on all p coefficient at ℓ=25. The trends for all dataset were similar, except for the dataset 500. It had high fluctuation and the EER raised rapidly.

6.1.3 Likelihood Ratio

The observation of LDA performance on Likelihood classifier can be seen in the figures 7 and 8. Figure 7 showed that the ℓ coefficients become flat at $\ell > s - 1$. As the proof in [3], a higher dimensionality $\ell > subject - 1$ did not contribute to the Likelihood ratio. So, the optimal area of Likelihood ratio was in the dimension $d \leq (p, s - 1)$. On the other hand, the optimal performance on p observation, as shown in figure 8, was in the range p=50 to 200 for dataset larger than 1000.

6.2 The Impact of Number of Images per Subject

Table 1 showed the impact of images per subject to recognition performance. The EER was growing down while the number of images per subject was increased for all classifiers. It means that the addition of number of images per subject will produce the better performance. However, the different of EER between them becomes smaller with double increase in the number of images. So, the more addition of images per subject will not give contribution significantly to recognition, but only give the risk of slower computation.

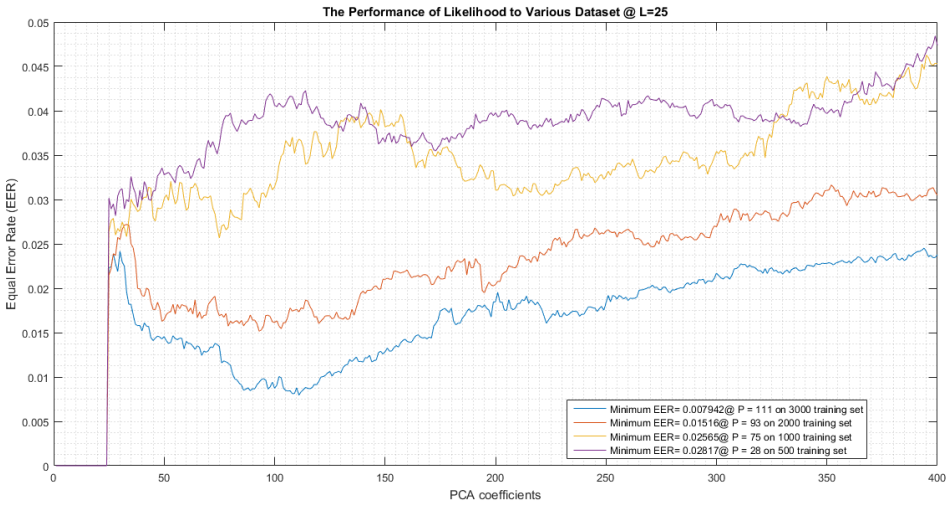


Figure 8: LDA coefficients using Likelihood Ratio on various dataset at ℓ=25

Table 1: Impact of Images per subject

Subjects	Images per	Equal Error Rate(EER)		
	Subject	Likelihood	Cosine	Euclidean
25	2	0.2890	0.1724	0.1946
25	5	0.1113	0.1512	0.1751
25	10	0.1019	0.1264	0.1610
25	20	0.0974	0.1224	0.1495

6.3 The Impact of Number of Images per subject and Number of Subjects trade off

Table 2: Number of images per subject and number of subjects trade off

Subjects	Images per	Minimum Equal Error Rate(EER)		
	Subject	Likelihood	Cosine	Euclidean
200	2	0.3567	0.0573	0.172
80	5	0.0969	0.0556	0.112
40	10	0.0858	0.0504	0.101
20	20	0.0988	0.0758	0.112

The analysis of recognition performance on ℓ coefficients at $p = 200$ was shown in the figures 9,10, and 11. The optimal performance on three classifiers indicated the same characteristics: the 200x2 training set produced the worst performance, the 80x5 and 20x20 training set produced the similar performance, and the 40x10 yielded the best performance. The minimum Equal Error Rate on three classifiers, as shown in the table 2, also indicated that the best combination was the training set from 40 subjects with 10 images per subject. So, both of the number of subjects and the number of images per subject give contribution to recognition.

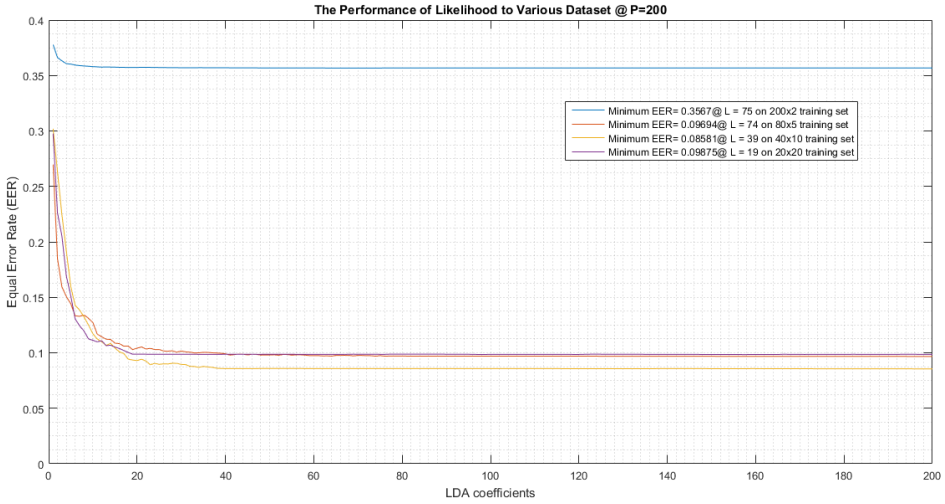


Figure 9: The impact of number of images per subject and number of subjects trade off on Likelihood classifier

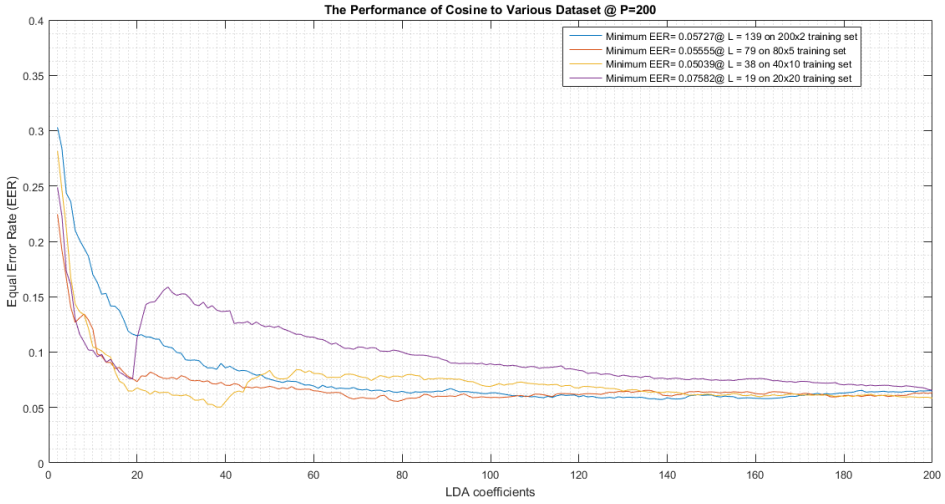


Figure 10: The impact of number of images per subject and number of subjects trade off on Cosine classifier

6.4 The Impact of Image Size

The results of this experiment were shown in the table 3. They showed that PCA-LDA can perform face recognition on very low resolution images up to 15x15 pixels. But, the performance will break down for the resolution less than 15x15 pixels. It was caused by the less information in the very small image. So, the PCA-LDA can not discriminate it well.

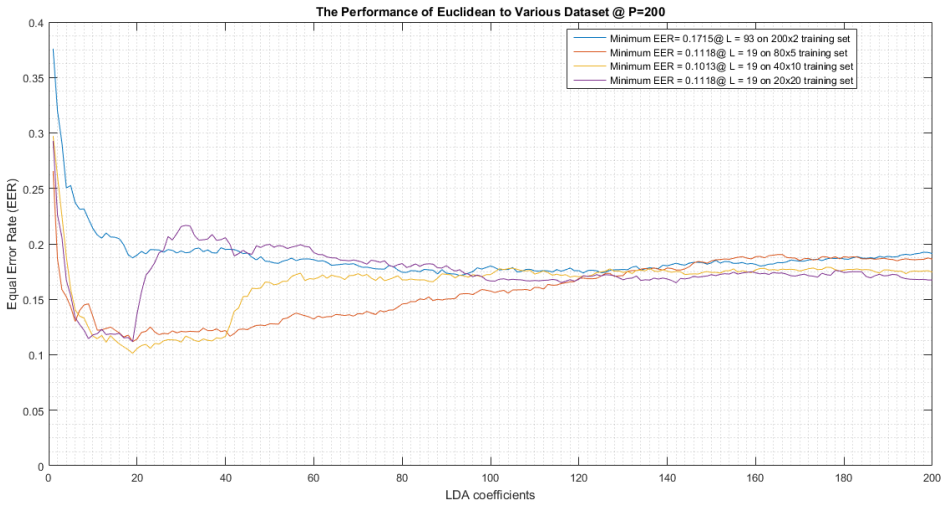


Figure 11: The impact of number of images per subject and number of subjects trade off on Euclidean classifier

Table 3: The Impact of Image Size

Image Size or	Equal Error Rate(EER)		
Resolution	Likelihood	Cosine	Euclidean
200 x 200	0.0178	0.0190	0.0324
100 x 100	0.0181	0.0198	0.0355
50 x 50	0.0176	0.0212	0.0295
40 x 40	0.0157	0.0209	0.0306
30 x 30	0.0166	0.0202	0.0324
25 x 25	0.0142	0.0177	0.0298
20 x 20	0.0163	0.0164	0.0372
15 x 15	0.0184	0.0165	0.0379
12 x 12	0.0259	0.0219	0.0443
10 x 10	0.1401	0.1470	0.1432

Conclusion

The behaviors of PCA-LDA on three classifiers were presented in this paper. For all classifier, the larger number of subjects will give the better performance. The number of images per subject also produced the performance as well as the number of subjects. The higher number of images per subject the better performance we obtained. Regarding the number of images per subject and number of subjects trade off, it indicated that both of them have contribution to recognition.

Regarding the p and ℓ coefficients, PCA-LDA has different behavior for each classifier. Firstly, Likelihood ratio had optimal performance at dimension $d \leq (p, s-1)$ and p in the range 50 to 200 for large training set, i.e. 2000 samples or more than 2000 samples. For the small training set,i.e. less than 2000, p coefficient was close to $s-1$ with $p > s-1$. The next classifiers, Euclidean and Cosine, have a bump at $\ell = s-1$ and form a valley at $\ell < s-1$ as optimal areas. The Cosine has optimal areas at ℓ= 10 to 100. The trend of cosine line is growing down after the bump and it is possible to

reach the minimum EER at higher ℓ coefficients. The Euclidean has different behavior with Cosine, the ℓ coefficients have optimal area at 10 to 50. ℓ were saturated after the bump for $\ell > s - 1$. Otherwise, the optimal range of p coefficients is p=80 to 120 for Euclidean and p=70 to 200 for Cosine classifier.

The PCA-LDA also has good performance on low resolution images. It can perform well up to 15x15 pixels resolution, but the performance will break down afterward.

Acknowledgements

The research described in this paper was supported by Research and Innovation in Science and Technology Project (RISET-Pro) of Ministry of Research, Technology, and Higher Education of Republic Indonesia (World Bank Loan No.8245-ID).

References

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991, pMID: 23964806. [Online]. Available: https://doi.org/10.1162/jocn.1991.3.1.71

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, Jul 1997.

[3] R. Veldhuis, A. Bazen, W. Booij, and A. Hendrikse, *Hand-Geometry Recognition Based on Contour Parameters*. SPIE – The Int. Society for Optical Engineering, 3 2005, pp. 344–353, imported from DIES.

[4] R. Veldhuis and A. Bazen, *One-to-template and ono-to-one verification in the single- and multi-user case*. Netherlands: Werkgemeenschap voor Informatie- en Communicatietheorie (WIC), 5 2005, pp. 39–46, imported from DIES.

[5] L. Spreeuwers, "Derivation of lda log likelihood ratio one-to-one classifier," vol. 2014, no. 1, p. 5, 2014.

[6] A. Sharma and K. K. Paliwal, "A two-stage linear discriminant analysis for face-recognition," *Pattern Recognition Letters*, vol. 33, no. 9, pp. 1157 – 1162, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865512000335

[7] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 531–542.

[8] N. Zanetti, A. Kumar, L. V. Huynh, and C. N. Teixidor, "Face recognition performance analysis: Impact of number of training samples and trade-off between number of individuals and images per individual," *University of Twente Students Journal of Biometrics and Computer Vision*, 2015. [Online]. Available: https://ojs.utwente.nl/index.php/UTSjBCV/article/view/4

[9] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," vol. 30, pp. 169–190,, 05 2017.

[10] F. Z. Chelali, A. Djeradi, and R. Djeradi, "Linear discriminant analysis for face recognition," in *2009 International Conference on Multimedia Computing and Systems*, April 2009, pp. 1–10.

[11] C. Zhang, Q. Ruan, and X. Pan, "Local fisher discriminant embedding for face recognition," in *2008 9th International Conference on Signal Processing*, Oct 2008, pp. 1660–1663.

[12] L. Spreeuwers, R. Veldhuis, S. Sultanali, and J. Diephuis, *Fixed FAR Vote Fusion of regional Facial Classifiers*. Gesellschaft fr Informatik, 9 2014, pp. 1–4.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[14] C. Tosik, A. Eleyan, and M. S. Salman, "Illumination invariant face recognition system," in *2013 21st Signal Processing and Communications Applications Conference (SIU)*, April 2013, pp. 1–4.

Capacity of the First-Order Low-Pass Channel with Power Constraint

Shokoufeh Mardani[1] and Jean-Paul Linnartz[1,2]
S.mardanikorani@tue.nl and J.p.linnartz@tue.nl
1. Electrical Engineering Dept., TU/e, Eindhoven, The Netherlands.
2. Signify - Philips Lighting Research, High Tech Campus, Eindhoven, The Netherlands.

***Abstract*—This paper bounds the performance of a uniform-power and a waterfilling algorithm for low-pass channels. It is shown that a visible light communication channel with double hetero-structure LEDs in their linear regime can be approximated by a first order low-pass channel. Closed form expressions are derived to relate the maximum achievable data rate (capacity) and the corresponding optimum bandwidth usage to the input total (modulation) power.**

***Index Terms*—LED, double hetero-structure, waterfilling, uniform power loading, capacity, optimum bandwidth, first order low pass channel, phosphor coated.**

I. Introduction

With the increasing pressure on the available radio spectrum, further accelerated by the Internet of Things, it is becoming increasingly important to optimally use the available bandwidth. Moreover optical wireless communication (OWC) also denoted as visible light communication (VLC) is being proposed as a new medium for wireless communication [1]. .

On the transmitter side of a VLC link, normally, light emitting diodes (LEDs) are used to convert the electrical driving current into light and to transmit information [2]. On the receiver side, an APD is needed to convert the received information into electrical current to be amplified, typically, by a transimpedance amplifier (TIA) [2].

It has been widely accepted that LEDs introduce a frequency dependent low-pass channel response. This low-pass behaviour is due to the LEDs' dynamic response, e.g. formulated by carriers rate equations in [3], parasitics of the wirings and also due to the low-pass response of APDs. To overcome this LED bandwidth limitation and to achieve high date rates OFDM modulation is normally employed in a VLC link. Yet, the need to optimize the data rate by adaptively assigning a specific power and also the number of bits depending on the channel properties is essential.

In this paper we use the term capacity (or maximum achievable rate or simply achievable rate) without necessarily implying Shannon capacity [4]. Shannon capacity gives the maximum achievable rate that can be achieved by making optimum choice for the power spectral composition (distribution of the power across the available bandwidth). Ideally, **waterfilling** is known as the solution to this optimization problem [5]. However, we also study the **uniform** assigning of the power to each frequency. We assume that a specific (modulation) power is available to be distributed across different frequencies (sub-carriers). In this work the maximum achievable rate and the optimum modulation bandwidth for the waterfilling and for the uniform power loading algorithm are derived analytically and compared. We will show that both the algorithms achieve almost the same performance in terms of the achievable rate for the given total (modulation) power. However, waterfilling requires more bandwidth.

The rest of the paper is organized as follows. The channel model is discussed in section II. Discussions on the achievable rate and the modulation power are given in section III. Waterfilling and uniform power loading algorithms are discussed in sections IV and V, respectively. Finally, comparisons and conclusions are presented in section VI.

II. Channel Model

Recently double hetero-structure LEDs have attracted attentions due to the improved external quantum efficiency. The dynamic behaviour of such LEDs can be described by the following equations [3]

$$\frac{dn(t)}{dt} = \frac{J(t)}{qt_w} - B_{rec}n(t)(p_0 + n(t)) \tag{1}$$

$$P_{opt}(t) = A_w t_w E_p B_{rec}[p_0 n(t) + n^2(t)] \tag{2}$$

where J is the injected current density, $q = 1.6 \times 10^{-19}$ is the elementary charge, E_p is a single photon energy, B_{rec} is radiative recombination constant, t_w is the LED thickness, A_w is the surface area, p_0 is the doping concentration in active layer, $n(t)$ is the injected carrier concentration and P_{opt} is the output optical power.

It is shown that LEDs, governed by the dynamic equations (1) and (2), exhibits non-linear behaviour. For linear applications which are dealing with amplitude modulated signals such as OFDM [6] or pulse amplitude modulation (PAM) [7] the LED should be employed in its linear regime.

Two find the channel model, in this paper, we consider three practical cases.

A. High doping concentration p_0

If the input injected current density J is low enough to always assume $p_0 \gg n(t)$, then the dynamic equations (1) and (2) can be simplified to

$$\frac{dn(t)}{dt} = \frac{J(t)}{qt_w} - B_{rec}p_0 n(t) \tag{3}$$

$$P_{opt}(t) = A_w t_w E_p B_{rec} p_0 n(t) \quad (4)$$

Then the output optical power is a linear function of the input current governed by a first order differential equation with time constant $\tau_c = \frac{1}{B_{rec} p_0}$. In the frequency domain this is equivalent to a first order low-pass response with a 3dB cut-off frequency $f_c = \frac{B_{rec} p_0}{2\pi}$.

B. Low modulation index: Small signal model

The modulated input current density J can be written as

$$J = J_0(1 + \alpha J_{ac}(t)) \quad (5)$$

where J_0 is the DC term (bias point), $J_{ac}(t)$ is the normalized modulated current density around the bias point with $< J_{ac}(t) >= 0$ and α is the modulation index. In this section J_0 can be very large and therefore the assumption $p_0 \gg n(t)$ might not be valid. However, we assume that α is very small, i.e. $\alpha \ll 1$. The corresponding carrier concentration $n(t)$ can be represented as

$$n(t) = n_0(1 + \beta n_{ac}(t)) \quad (6)$$

where we have $\beta \ll 1$. Applying (5) and (6) into (1), we have

$$\begin{aligned} \beta n_0 \frac{dn_{ac}(t)}{dt} =& \frac{J_0(1 + \alpha J_{ac}(t))}{qt_w} - B_{rec} p_0 n_0 (1 + \beta n_{ac}(t)) \\ & - B_{rec} n_0^2 \left(1 + \beta n_{ac}(t)\right)^2 \end{aligned} \quad (7)$$

which for $\beta \ll 1$ can be simplified to

$$\begin{aligned} \beta n_0 \frac{dn_{ac}(t)}{dt} + \beta B_{rec} n_0 (p_0 + 2n_0) n_{ac}(t) = \\ \frac{J_0 \alpha}{qt_w} J_{ac}(t) + \frac{J_0}{qt_w} - B_{rec} p_0 n_0 - B_{rec} n_0^2 \end{aligned} \quad (8)$$

The last three terms in (8) is the bias point equilibrium and therefore vanishes. This further simplifies (8) into

$$\frac{dn_{ac}(t)}{dt} + B_{rec}(p_0 + 2n_0) n_{ac}(t) = \frac{J_0 \alpha}{\beta n_0 q t_w} J_{ac}(t) \quad (9)$$

which is a first order differential equation with a time constant

$$\tau_c = \frac{1}{B_{rec}(p_0 + 2n_0)} \quad (10)$$

To find n_0, we use the bias point equilibrium $\frac{J_0}{qt_w} - B_{rec} p_0 n_0 - B_{rec} n_0^2 = 0$, which results in

$$n_0 = \frac{-p_0}{2} + \sqrt{\frac{p_0^2}{4} + \frac{J_0}{B_{rec} q t_w}} \quad (11)$$

Applying (11) into (10) results in

$$\tau_c = \frac{1}{\sqrt{(B_{rec} p_0)^2 + \frac{4 B_{rec} J_0}{qt_w}}} \quad (12)$$

which gives the channel time constant as a function of the LED parameters and the bias current density J_0. The channel response then follows a first order low-pass model with 3dB cut-off frequency

$$f_c = \frac{1}{2\pi} \sqrt{(B_{rec} p_0)^2 + \frac{4 B_{rec} J_0}{qt_w}} \quad (13)$$

C. High speed phosphor coated blue LEDs

For joint illumination and communication (JTC) systems, white LEDs are required [8]. Conventionally, a phosphor coated blue chip [2], [11] is used to convert a fraction α of the emitted blue photons (by the blue chip) into excited phosphor particles subsequently re-emitted as yellow photons. The white light is then constructed by the combination of the remaining blue and the converted yellow photons.

Let's assume that the blue chip is used in its linear regime and is fast enough not to affect the channel frequency response in the modulation bandwidth. Also, let's assume that the flow of the incoming blue photons per unit of time is denoted as $\phi_B(t)$. For a decay time constant τ, the outflow of the yellow photons $\phi_Y(t)$ can be expressed as $\phi_Y(t) = m(t)/\tau$ where $m(t)$ is the density of the phosphor particles in excited state.

Using this simplified model, the dynamics of $m(t)$ are governed by

$$\frac{dm(t)}{dt} = \alpha \phi_B(t) - \frac{m(t)}{\tau} \quad (14)$$

thus, in frequency domain

$$j2\pi f \tau \Phi_Y(f) = \alpha \Phi_B(f) - \Phi_Y(f) \quad (15)$$

This explains a first-order low pass behaviour

$$H_Y(f) = \frac{\Phi_Y(f)}{\Phi_B(f)} = \frac{\alpha}{1 + j2\pi f \tau} \quad (16)$$

The fraction $1 - \alpha$ passes through the phosphor without any conversion [2]. The channel transfer function can then be written as

$$H(f) = (1 - \alpha) + \frac{\alpha}{1 + j2\pi f \tau} = \frac{1 + j(1 - \alpha)2\pi f \tau}{1 + j2\pi f \tau} \quad (17)$$

where $\frac{1}{2\pi\tau} = f_c$ is the 3dB cut-off frequency of the phosphor. For the case of warm white LEDs α approaches 1 which results in a high frequency for the zero of the transfer function in (17). Also, high frequency poles due to the parasitics and also due to the APD cancel out the effect of this high frequency zero. Therefore, for phosphor coated blue LEDs, normally, the first order low-pass response of the phosphor dominates the channel model.

As a conclusion, this theoretical paper addresses the first order low-pass channel model which for this channel model three different scenarios were discussed. We use the following model for the channel frequency response

$$H(f) = H_0 \left(1 + \frac{f^2}{f_c^2}\right)^{-\frac{1}{2}} \quad (18)$$

where H_0 is the low frequency channel gain which for simplicity we assume $H_0 = 1$ and f_c is the 3dB cut-off frequency of the channel. This model is a reasonable model for illumination LEDs used to communicate via visible light, yet our results also hold for other instantiations of a low-pass channel.

We assume that the dominant noise is the shot noise from the ambient light or the thermal noise caused by the TIA that is following the photo-detector. Both of these noise sources can be accurately modelled as additive white Gaussian noise (AWGN) which is independent from the transmitted signal

[12]. The Power Spectral Density (PSD) of the noise is represented by N_0 (Watts/Hz).

III. Achievable Rate

The achievable rate of the OFDM depends on the power allocated to the various sub-carriers. We can maximize the rate by choosing an appropriate power allocation subject to the limited total modulation power P_{tot}. The achievable rate can be written as

$$\max \int_f \ln\left(1 + P(f)\frac{|H(f)|^2}{N_0}\right) df \tag{19}$$

where $P(f)$ is the allocated power to frequency f. The constraint on the total allocated power implies that

$$\int_f P(f)df \leq P_{tot} \tag{20}$$

For a VLC link, a constraint on the transmitted (modulation) power can be misleading as it has been discussed that in VLC links the modulation power is available for free. In [8]–[10], it is shown that modulating the LED requires an extra power on top of the illumination power. This extra power is derived to be proportional to the second moment of the modulation current. Then, the constraint, given in (20), is equivalent to have a constraint on the extra power, beyond illumination power, needed to AC modulate the LED current. This makes the communication channel a power limited channel which then justifies the use of optimization principles such as waterfilling. Moreover, this justifies the use of Gaussian signal distribution as in DCO-OFDM and that the capacity of the channel is the sum (or integral) of the capacities per sub-band.

In the following we consider two specific power allocations and derive the achievable rate and the optimum modulation bandwidth for the given total modulation power.

A. Waterfilling

The generic solution that maximizes the achievable rate is known as the waterfilling. [5]. However, for the specific low-pass channel, given by (18), we derive elegant expressions. Our approach is to derive both the total modulation power P_{tot} and the achievable rate as a function of the optimum modulation bandwidth.

The generic solution to maximize the achievable rate is as following [5]

$$P(f) = \left[v - \frac{N_0}{|H(f)|^2}\right]^+ \tag{21}$$

where v is a constant. There is no guarantee that $P(f)$ arises from a choice of v in (21) is non-negative in general. We choose $v = \frac{N_0}{|H(f_{\max_w})|^2}$ to make $P(f_{\max_w}) = 0$. Then, $f_{\max_w}$ is the maximum frequency that the assigned power is non-zero. Then for $f > f_{\max_w}$ we have $P(f) = 0$; $f_{\max_w}$ is denoted as the optimum modulation bandwidth for the waterfilling. Using the channel model, given by (18), $P(f)$ can be written as

$$P(f) = \frac{N_0}{f_c^2}\left(f_{\max_w}^2 - f^2\right) \tag{22}$$

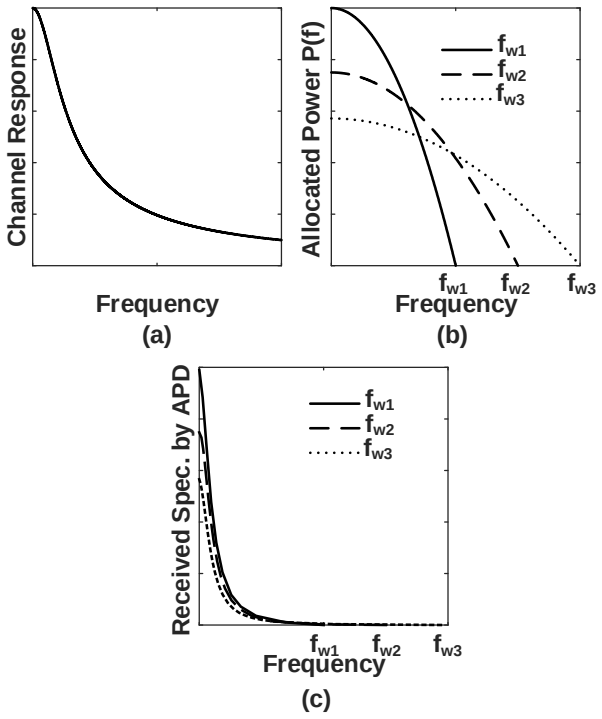


Fig. 1. Illustration of the waterfilling algorithms for three different values of $f_{\max_w}$, $f_{w3} > f_{w2} > f_{w1}$. (a) Channel response, (b) allocated power $P(f)$ and (c) the received power spectrum by the APD.

For three different values of $f_{\max_w}$, $f_{w3} > f_{w2} > f_{w1}$, the normalized channel response, the normalized allocated power and the normalized received power spectrum by the APD (which is proportional to $P(f)|H(f)|^2$) are illustrated in Fig. 1. It can be seen that the allocated power is a decreasing function of the frequency and that for a higher $f_{\max_w}$, under the constraint given by (20), the allocated power at lower frequencies decreases.

Applying $P(f)$, given by (22), in (19) results in the achievable rate for the waterfilling, R_w, as follows

$$R_w = \int_0^{f_{\max_w}} \ln\left(\frac{f_{\max_w}^2 + f_c^2}{f^2 + f_c^2}\right) df = f_{\max_w} \ln\left(f_{\max_w}^2 + f_c^2\right) - \int_0^{f_{\max_w}} \ln\left(f^2 + f_c^2\right) df \tag{23}$$

Using the integration by parts technique, it can be shown that

$$\int \ln\left(x^2 + a^2\right) df = x \ln\left(x^2 + a^2\right) - 2x + 2a \tan^{-1}\left(\frac{x}{a}\right) \tag{24}$$

Therefore, equation (23) can be simplified to

$$\frac{R_w}{f_c} = 2\left(\frac{f_{\max_w}}{f_c}\right) - 2\tan^{-1}\left(\frac{f_{\max_w}}{f_c}\right) \tag{25}$$

which relates the achievable rate of the waterfilling algorithm to the optimum modulation bandwidth. To find a relation between the total allocated power P_{tot} and the optimum modulation bandwidth $f_{\max_w}$, we apply $P(f)$ (given by (22)) into the constraint on the total power (given by (20)). This results in

$$P_{tot} = \int_0^{f_{\max_w}} \frac{N_0}{f_c^2} \left(f_{\max_w}^2 - f^2\right) df \tag{26}$$

which can be simplified to

$$\frac{P_{tot}}{N_0 f_c} = \frac{2}{3} \left(\frac{f_{\max_w}}{f_c}\right)^3 \tag{27}$$

The unique relation between the maximum achievable rate R_w and the total allocated power P_{tot} can be readily obtained from (25) and (27) as [13]

$$\frac{R_w}{f_c} = 2\left(\frac{3}{2}\frac{P_{tot}}{N_0 f_c}\right)^{(1/3)} - 2\tan^{-1}\left(\left(\frac{3}{2}\frac{P_{tot}}{N_0 f_c}\right)^{(1/3)}\right) \tag{28}$$

For the given total transmitted power, the $f_{\max_w}$ is obtained from (27) and then the achievable rate from (28).

IV. Uniform power loading

The simplest way to load the sub-carriers is loading all of them with a same power. Assuming that the power is uniformly distributed across the frequency range $[0, f_{\max_u}]$, we have

$$P(f) = \frac{P_{tot}}{f_{\max_u}} \tag{29}$$

where $f_{\max_u}$ is the optimum modulation BW for the uniform loading strategy. Note that for $f > f_{\max_u}$ we have $P(f) = 0$. Fig. 2 illustrates the uniform power loading algorithm for three different values of $f_{\max_u}$, $f_{u3} > f_{u2} > f_{u1}$.

From (19) and (29), the achievable rate can be written as

$$\begin{aligned} R_u &= \int_0^{f_{\max_u}} \ln\left(1 + \frac{P_{tot}}{N_0 f_{\max_u}} \frac{1}{1+\left(\frac{f}{f_c}\right)^2}\right) df \\ &= \int_0^{f_{\max_u}} \ln\left(f^2 + f_c^2(1 + \frac{P_{tot}}{N_0 f_{\max_u}})\right) df \\ &\quad - \int_0^{f_{\max_u}} \ln\left(f^2 + f_c^2\right) df \end{aligned} \tag{30}$$

Using the integral equation in (24), equation (30) can be simplified to

$$\begin{aligned} R_u &= f_{\max_u} \ln\left(\frac{f_{\max_u}^2 + f_c^2(1+\frac{P_{tot}}{N_0 f_{\max_u}})}{f_{\max_u}^2 + f_c^2}\right) \\ &\quad + 2f_c\sqrt{1+\frac{P_{tot}}{N_0 f_{\max_u}}}\tan^{-1}\left(\frac{f_{\max_u}}{f_c\sqrt{1+\frac{P_{tot}}{N_0 f_{\max_u}}}}\right) \\ &\quad - 2f_c\tan^{-1}\left(\frac{f_{\max_u}}{f_c}\right) \end{aligned} \tag{31}$$

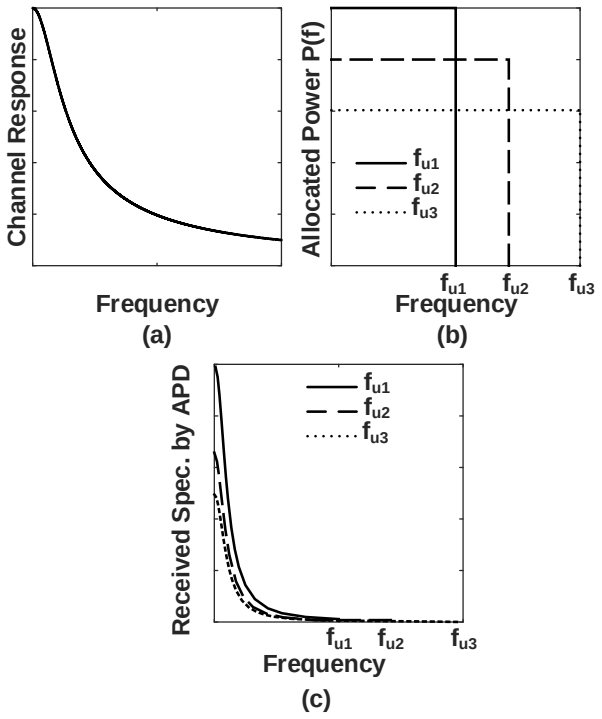


Fig. 2. Illustration of the uniform power loading algorithms for three different values of $f_{\max_u}$, $f_{u3} > f_{u2} > f_{u1}$. (a) Channel response, (b) allocated power $P(f)$ and (c) the received power spectrum by the APD.

To find a relation between the total allocated power P_{tot} and the optimum bandwidth $f_{\max_u}$, one can use $\frac{dR_u}{df_{\max_u}} = 0$. This results in the optimum bandwidth $f_{\max_u}$ for the given P_{tot} that maximizes the achievable rate R_u.

V. Comparison and conclusion

Fig. 3 compares the two algorithms in plots that show the achievable rate and the optimum modulation bandwidth as a function of $\frac{P_{tot}}{N_0 f_c}$. It can be seen that both the algorithms achieve almost the same performance in terms of the achievable rate. However, waterfilling demands more bandwidth (almost 25%). To compare the rate performance, we define the relative rate difference as

$$\frac{\Delta R_{wu}}{R_w}(\%) = 100 \times \frac{R_w - R_u}{R_w} \tag{32}$$

The relative rate difference of the two algorithms is shown in Fig. 4. It can be seen that the relative rate difference between the two algorithms is less than 2% (relatively).

In this paper we derived analytical expressions for the low pass channel. These formulas are highly relevant to assess the capacity and the optimum modulation bandwidth of Visible Light Communication channels. The accurate prediction of the occupied bandwidth and corresponding rate are useful in practical applications to e.g. specify the sampling rate and to design filters.

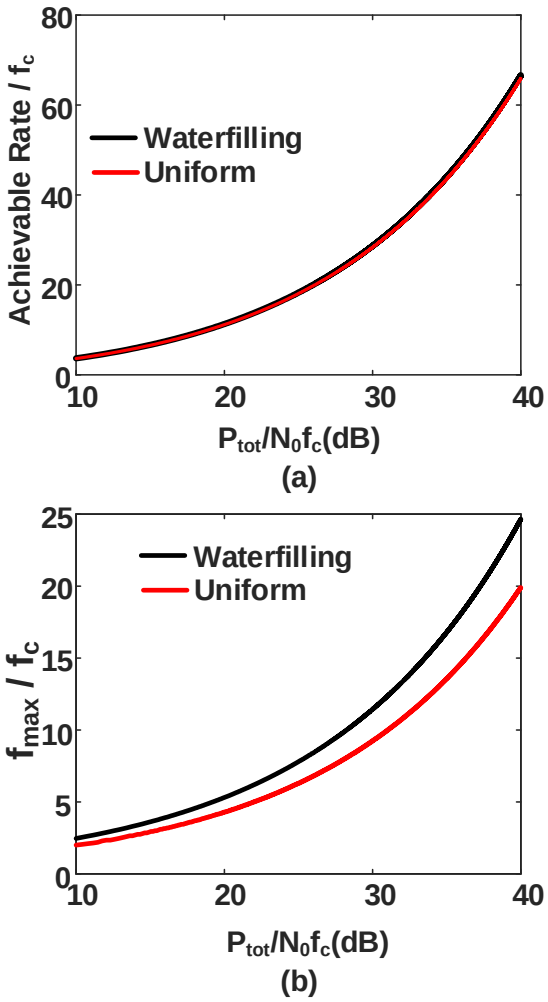


Fig. 3. (a) Achievable rate and (b) optimum modulation BW as a function of $\frac{P_{tot}}{N_0 f_c}$.

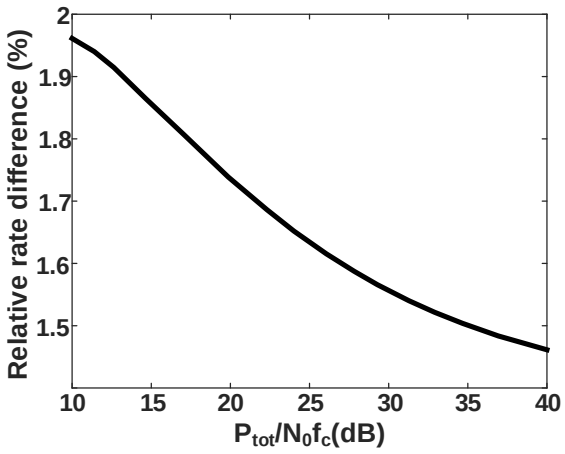


Fig. 4. Relative rate difference of the waterfilling and uniform power loading algorithm as a function of $\frac{P_{tot}}{N_0 f_c}$.

REFERENCES

[1] Visible Light Communication Potential Solution to the Global Wireless Spectrum Shortage GBI Research, Tech. Rep., 2011.

[2] S. Mardani, A. Khalid, F. M. J. Willems and J. P. Linnartz, "Effect of Blue Filter on the SNR and Data Rate for Indoor Visible Light Communication System," 2017 European Conference on Optical Communication (ECOC), Gothenburg, Sweden, 2017, pp. 1-3.

[3] R. Windisch, A. Knobloch, M. Kuijk, C. Rooman, B. Dutta, P. Kiesel, G. Borghs, G. Dohler, and P. Heremans, Large-signal-modulation of high-efficiency light-emitting diodes for optical communication, IEEE journal of quantum electronics, vol. 36, no. 12, pp. 14451453, 2000.

[4] C. E. Shannon, "A mathematical theory of communication," in The Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, July 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x

[5] R. G. Gallager, Information Theory and Reliable Communication. New York: Wiley, 1968.

[6] J. Armstrong, OFDM for Optical Communications, IEEE/OSA Journal on Lightwave Technology (IEEE/OSA JLT), vol. 27, no. 3, pp. 189204, Feb. 2009.

[7] Stepniak Grzegorz, L. Maksymiuk, J. Siuzdak, "1.1 GBIT/S white lighting LED based visible light link with pulse amplitude modulation and Volterra DFE equalization", Microwave and Optical Technology Letters, vol. 7, no. 57, pp. 1620-1622, 2015.

[8] A. Tsiatmas, F. M. Willems, J. P. M. Linnartz, S. Baggen and J. W. Bergmans, "Joint illumination and visible-Light Communication systems: Data rates and extra power consumption," 2015 IEEE International Conference on Communication Workshop (ICCW), London, 2015, pp. 1380-1386. doi: 10.1109/ICCW.2015.7247371

[9] J. P. M. G. Linnartz, "Wireless optical communication in illumination systems," 2016 IEEE Photonics Society Summer Topical Meeting Series (SUM), Newport Beach, CA, 2016, pp. 104-107. doi: 10.1109/PHOSST.2016.7548764

[10] Xiong Deng, Yan Wu, A. M. Khalid, Xi Long, and Jean-Paul M. G. Linnartz, "LED power consumption in joint illumination and communication system," Opt. Express 25, 18990-19003 (2017).

[11] A. M. Khalid, G. Cossu, R. Corsini, P. Choudhury and E. Ciaramella, "1-Gb/s Transmission Over a Phosphorescent White LED by Using Rate-Adaptive Discrete Multitone Modulation," in IEEE Photonics Journal, vol. 4, no. 5, pp. 1465-1473, Oct. 2012. doi: 10.1109/JPHOT.2012.2210397

[12] J. M. Kahn and J. R. Barry, "Wireless infrared communications," in Proceedings of the IEEE, vol. 85, no. 2, pp. 265-298, Feb 1997.

[13] J. Lee, Discrete multitone modulation for short-range optical communications, Eindhoven: Technische Universiteit Eindhoven, 2009.

Enabling Distributed Transmit Diversity by Wireless Synchronization for IEEE 802.11p

L. M. A. van Meurs A. G. C. Koppelaar A. Filippi M. van Splunter
NXP Semiconductors
Eindhoven, The Netherlands
lars.van.meurs@nxp.com

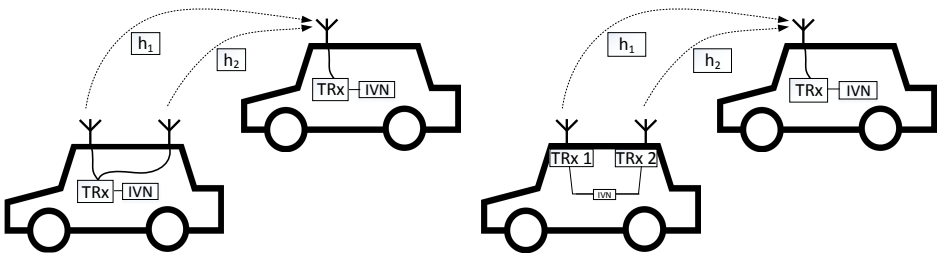
Abstract

In this paper the challenges that distributed radios with independent oscillators bring for diversity transmissions in a V2X communication system employing IEEE 802.11p are studied. The work focuses specifically to dual antenna transmit diversity using cyclic delay diversity. Synchronization requirements for operating without loss of performance in a distributed transmit diversity system are determined and methods to do the synchronization wirelessly are investigated. The proposed method is implemented on an IEEE 802.11p hardware platform and the resulting performance is evaluated. This work is the first step towards a distributed architecture for a multiple antenna wireless communication system.

1 Introduction

Vehicle-to-vehicle and vehicle-to-infrastructure (V2X) communication enables vehicles to communicate to other vehicles and to the infrastructure. This will make transportation systems safer, more efficient and more enjoyable. The V2X technology that is ready for deployment today is IEEE 802.11p based [1]. To meet the increasing performance requirements, the automotive industry is continuously investigating ways to improve the robustness of V2X systems under harsh automotive conditions. One way to do this is to realize diversity by increasing the number of receive and/or transmit antennas.

For the best performance, these antennas should be located at well-separated locations on the vehicle. In the current system architectures, the antennas are connected by expensive RF cables and compensators to a single transceiver containing multiple radios. From a system cost perspective, the transceiver should be located close to the antenna. In that way, the RF signals can be processed close to the antenna and a less expensive, and potentially shared, interface can be used to exchange the information over the in-vehicle network (IVN). For multiple antenna systems, this requires splitting of a multi-radio transceiver into multiple single radio transceivers located closely to each antenna. This distributed architecture will scale much better with the increasing number of antennas and is more modular.



(a) Conventional (b) Distributed

Figure 1: System architectures

The distributed architecture poses a challenge. Due to the partition of radios over multiple well-separated transceivers, it is more difficult to keep these radios synchronized to each other. Each radio has its own independent oscillator, and therefore has a slightly different time/frequency reference. Diversity techniques are in general relying on a common reference for the transmission of signals. Known solutions to synchronize time and frequency use a wired synchronization interconnect (costly), depend on GPS (not always available and costly) or require feedback from the receiving party [2] (not supported by the standard). Achieving receive diversity in a distributed architecture has already been discussed [3].

The goal of this work is to develop a wireless synchronization solution for distributed transmit diversity for IEEE 802.11p [4] without the need for additional hardware and negligible loss of performance. This work is the first step towards a distributed architecture for a multiple antenna wireless communication system.

IEEE 802.11p technology uses coded orthogonal frequency-division multiplexing (C-OFDM) modulation with a bandwidth of 10 MHz in the 5.9 GHz band and several modulation and coding schemes (MCSs). Although not standardized in IEEE 802.11p, dual antenna transmit diversity is used to improve the performance of V2X communication. This transmit diversity is realized by means of a mutual cyclic delay on the OFDM symbols according to spatial expansion as in IEEE 802.11n and also known as cyclic delay diversity (CDD) [5]. CDD is a form of transparent transmit diversity, i.e. it does not require modification of an IEEE 802.11p single antenna receiver.

This paper is organized as follows. In Section 2 the challenges of distributed transmit diversity are studied. In Section 3 a wireless synchronization solution to overcome these challenges is discussed. The performance of the solution is reported in Section 4 and conclusions are provided in Section 5.

2 Challenges of Distributed Transmitters

This section describes the challenges that a distributed transmitter architecture introduces for receiving a CDD signal (Section 2.1) and the impact of oscillator drift (Section 2.2). The main difference between a conventional and the distributed architecture is, that each distributed transmitter has its own independent oscillator as illustrated in Figure 2. It also shows that the carrier and sampling frequencies are derived from the same source, as dictated by the IEEE 802.11p standard.

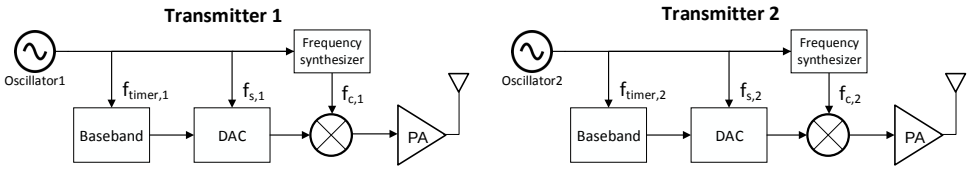


Figure 2: Independent oscillators and frequencies in a distributed architecture

Independent oscillators will generate slightly different frequency reference signals. This will cause carrier frequency offset (CFO), transmission time offset (TO) because of baseband timer frequency offset and sampling frequency offset (SFO) between the transmitted frames of the dual diversity transmissions. They also introduce a random phase shift between transmitted signals. This is not a problem and desired anyway [6]. The payload to be transmitted is provided to both transmitters using the IVN.

2.1 Impact on Receiver

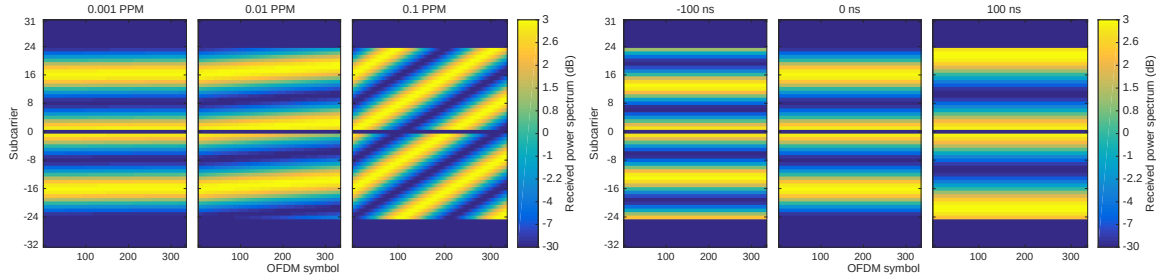
The IEEE 802.11p standard is not developed for a distributed architecture. This means that receivers are not designed to cope with the aforementioned CFO, SFO and TO.

We describe an OFDM-based digital communication system by defining $s_n(t)$ as the n^{th} transmitted OFDM symbol in time domain and $S_n(k)$ as the n^{th} transmitted OFDM symbol in frequency domain, where t is the time variable and $k = 0, 1 \ldots K-1$ is the subcarrier index, where K is the total number of subcarriers. In a similar way, the received OFDM symbols are represented by $r_n(t)$ and $R_n(k)$, the channel by $h(t)$ and $H(k)$ and the noise by $n_n(t)$ and $N_n(k)$. Received OFDM symbols can be expressed as follows when CFO is introduced between the transmitters

$$r_n(t) = \frac{1}{2}\sqrt{2} \cdot h(t) \cdot \Bigg(s_n(t + \frac{1}{2}T_{\text{D}}) \exp\left(j\frac{1}{2}\phi + j\pi\Delta f_c t\right) + s_n(t - \frac{1}{2}T_{\text{D}}) \exp\left(-j\frac{1}{2}\phi - j\pi\Delta f_c t\right) \Bigg) + n_n(t) \tag{1}$$

$$R_n(k) = \sqrt{2} \cdot H(k) \underbrace{\left(\alpha \cdot S_n(k) + N_{\text{ici}}(k)\right)}_{\text{ICI}} \cos\left(\pi \Bigg(\underbrace{\frac{k}{K} f_s T_{\text{D}}}_{\text{Freq. select.}} + \underbrace{\Delta f_c n T_s}_{\text{Fading}} \Bigg) + \frac{1}{2}\phi \right) + N_n(k) \tag{2}$$

where T_{D} is the cyclic delay, ϕ the phase difference at frame-start, f_s the sampling frequency, Δf_c the CFO, and α / N_{ICI} the inter-carrier-interference (ICI). This result shows that the received frame is an amplitude modulated (AM) version of the transmitted frame with a modulation frequency that is half the CFO (frequency selectivity is due to CDD). This AM is perceived as (self-induced) channel fading by the receiver. Besides this, ICI is introduced because the subcarriers of each transmitter are not perfectly aligned anymore. In order to perform well in mobility scenarios (e.g. on the highway) IEEE 802.11p receivers should be designed to withstand channel fading by employing channel tracking and equalization [7]. Therefore, the receiver's ability to handle fading, would depend on its channel tracking capabilities. In Figure 3a the perceived frequency response on an AWGN channel for a distributed CDD transmitter with several CFO values is shown.



(a) CFO (b) TO

Figure 3: Perceived frequency response of distributed transmitter for AWGN channel

In a similar fashion it can be shown that SFO also results in ICI and fading. The impact of SFO can be accounted for by a similar expression as the expression for CFO (1,2) by replacing Δf_c with $k \cdot \Delta f_s$. Since $k \cdot \Delta f_s << \Delta f_c$ for IEEE 802.11p the impact of SFO will be much smaller than that of CFO. Next to this, the SFO will cause a decrease in the effective guard interval, as the symbol timing of the transmitters will relatively change over time and will thus spread the signal. However, this effect is negligible for typical IEEE 802.11p packet lengths and allowed frequency offsets.

Received OFDM symbols can be expressed as follows when TO is introduced

$$r_n(t) = \frac{1}{2}\sqrt{2} \cdot h(t) \cdot \Big(s_n(t + \frac{1}{2}T_\mathrm{D} + \frac{1}{2}\Delta t) \exp\left(j\frac{1}{2}\phi\right) + s_n(t - \frac{1}{2}T_\mathrm{D} - \frac{1}{2}\Delta t) \exp\left(-j\frac{1}{2}\phi_n\right) \Big) + n_n(t) \tag{3}$$

$$R_n(k) = \sqrt{2} \cdot H(k) \cdot S_n(k) \cos\left(\underbrace{j\pi \frac{k}{K} f_s(T_\mathrm{D} + \Delta t)}_{\text{Frequency selectivity}} + \frac{1}{2}\phi \right) + N_n(k) \tag{4}$$

where Δt is the time offset between the two transmitted frames. This result shows that TO changes the frequency selectivity introduced by CDD. If the TO is in line with the cyclic delay, the frequency selectivity will increase but the interoperability with legacy receivers will decrease [8]. If the time offset is opposite to the cyclic delay, the frequency selectivity will decrease and therefore the resilience against wide channel fades. In Figure 3b the perceived frequency response of an AWGN channel for a distributed CDD transmitter and several TOs is shown. Another consequence of the TO is a decrease of the effective cyclic prefix.

The Physical Layer (PHY) performance of an IEEE 802.11p receiver is simulated while CFO, SFO or TO is applied to a dual diversity transmitter. This is done for all MCSs, for receivers with different channel tracking capabilities and both the IEEE 802.11 channel model D [9] and the AWGN channel. Figure 4 shows the results of the CFO simulations for MCS7 in Channel D. Table 1 summaries all results by indicating the maximum allowable offset for <0.2 dB impact on the BER performance.

Ch. tracking latency [μs]	**CFO [PPM]**	**TO [ns]**	**SFO [PPM]**
0	0.25	50	20*
8	0.1	50	20*
16	0.05	50	20*
40	0.02	50	20*
∞	0.001*	50	5*

Table 1: Maximum allowable CFO, TO and SFO for <0.2 dB impact on the BER performance of of the receiver for all MCS

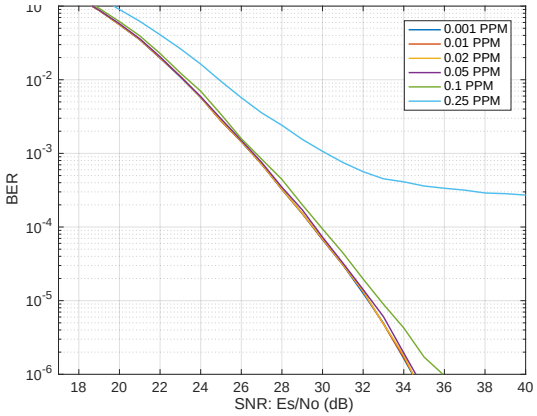


Figure 4: BER performance for distributed CFO on MCS7 in Channel D

2.2 Oscillator Offset & Drift

The IEEE 802.11p standard allows transceivers to have a maximum oscillator frequency offset of 20 PPM, which means that distributed transceivers could have a 40 PPM relative offset. Due to frequency instability of the oscillators, this offset will not be constant over time. This so-called drift is mainly caused by variations in temperature. Although temperature controlled crystal oscillators (TCXOs) can be used, frequency instability is still an issue. We have approximated the worst-case oscillator drift rate

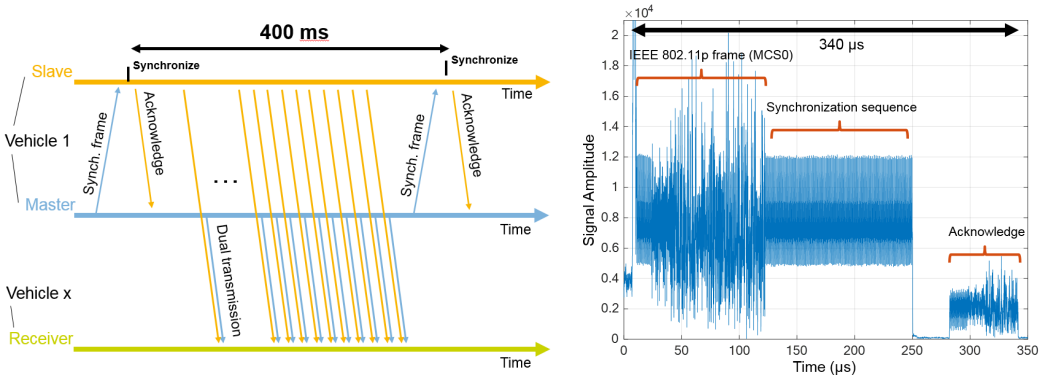
*Assuming 1000 bytes IEEE 802.11p MCS0 frames. Shorter packets allow a higher limit.

by accurately tracking the oscillator offset over time with several samples of a NXP RoadLINK equipped IEEE 802.11p based V2X solution, which uses a TCXO.

Two causes for oscillator drift were considered and simulated in these measurements; temperature variation due to self-heating and external temperature variations. The first is strongly depended on the system design, e.g. the thermal isolation between dissipating components and the TCXO, and the TCXO choice. A maximum drift rate of 0.05 PPM/s was observed when full power transmission started at an initial ambient temperature of $-40\,°C$. For the oscillator drift due to ambient temperature variations we assume a maximum temperature variation of $1\,°C/s$ at the TCXO. We have applied this temperature gradient in both directions in a -40 to $90\,°C$ temperature range and observed a maximum drift of approximately 0.05 PPM/s (in the 80-90 °C temperature range). From these experiments we conclude a worst-case relative oscillator drift of 0.1 PPM/s (since they can drift oppositely) between the transmitters' oscillators.

3 Synchronization Solution

For the solution design we assume that receivers employ channel tracking with a latency of at most $16\,\mu s$. This results in a maximum CFO, TO and SFO requirement of respectively 0.05 PPM, 50 ns and 20 PPM (Table 1), which should be maintained even when oscillators drift relatively with a maximum of 0.1 PPM/s (Section 2.2). This section describes a wireless synchronization procedure developed for and implemented on an NXP RoadLINK equipped V2X modem.



(a) Timeline (b) Synchronization frame exchange

Figure 5: Wireless synchronization solution

We propose to synchronize the two transceivers by periodically transmitting a synchronization frame from one distributed transceiver (master) to the other (slave) as illustrated in Figure 5a. The slave will use the synchronization frame to determine the time and frequency offsets, correct for them and respond to the master with an acknowledgment frame. Once synchronized the master and slave can perform diversity transmissions together. Periodic synchronization is needed to track the oscillator drift. We choose to estimate and correct the frequency offsets with a 0.01 PPM accuracy every 400 ms in order to remain within the offset limits with a worst-case relative oscillator drift of 0.01 PPM/s. We recognize that this could be improved by estimating oscillator drift and predicting frequency offset.

The synchronization frame should contain a repeating or known sequence so that the slave can estimate the CFO by performing an auto- or cross-correlation on this received sequence [10]. In our implementation the auto-correlation method is chosen. TO estimation is done by estimating the time-of-transmission of the synchronization frame at the master ($t_{sync-tx}$), the time-of-reception at the slave ($t_{sync-rx}$) and the

time-of-flight (TOF). The TO equals

$$\Delta t = t_{sync-tx} + \text{TOF} - t_{sync-rx} \quad , \quad \text{TOF} \approx \frac{t_{ack-rx} - t_{sync-tx} - T_{\text{sync}} - T_{\text{SIFS}}}{2} \tag{5}$$

where t_{ack-rx} is the time-of-arrival of the acknowledgment frame at the master, T_{sync} the synchronization frame duration and T_{SIFS} the IEEE 802.11p short interframe spacing. The RoadLINK based V2X modem has provisions to estimate the time-of-transmission and time-of-reception of IEEE 802.11p frames. Therefore, we use an IEEE 802.11p frame, appended by the frequency synchronization sequence, as the synchronization frame and a standard acknowledgment frame as shown in Figure 5b. The $t_{sync-tx}$ can be predicted before transmission takes place and will be included in the payload of the IEEE 802.11p-part of the synchronization frame, so that the receiver will have access to both $t_{sync-tx}$ and $t_{sync-rx}$. The master will estimate the TOF (5) every synchronization interval. Since the master and slave have a fixed location on the vehicle, we assume that the TOF is constant and we use filtering techniques to determine a more accurate estimate. We include this estimation in the payload of the (next) synchronization frame so that the slave can calculate the TO. We have experimentally found that when no drift is present, the RoadLINK modem can estimate the TO with an accuracy of ± 35 ns. When worst-case drift is present, this would result in a maximum TO of

$$\text{TO}_{\text{max}} = 35 \text{ ns} + 0.01 \text{ PPM} \cdot 0.4\text{s} + \frac{1}{2} 0.1 \text{ PPM/s} \cdot (0.4\text{s})^2 = 47 \text{ ns} \tag{6}$$

when the slave baseband timer clock skew is minimized to 0.01 PPM as described hereafter. Correction of the CFO can be done by performing a frequency shift to the digital baseband samples (complex multiplication with $e^{j2\pi\Delta f t}$) of each outgoing frame at the slave or by tuning the carrier frequency synthesizers. We have chosen the latter. Correction of the SFO can be done by resampling (with $\Delta f_c \cdot f_s / f_c$) at the slave. We have not implemented SFO correction, because the used oscillators already meet the SFO offset requirement. Correction of the TO will be done by offsetting the slave baseband time by the TO. Furthermore, a frequency offset (of $\Delta f_c \cdot f_{\text{timer}} / f_c$) is applied to the baseband timer to minimize the clock skew.

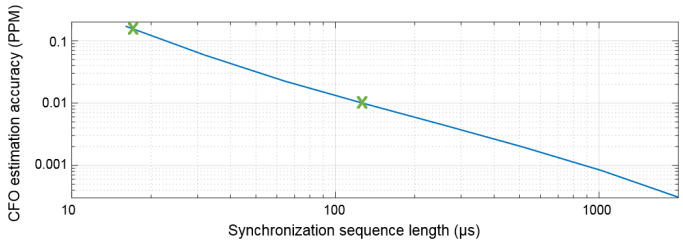


Figure 6: Simulation of CFO estimator performance for Channel D, 30 dB SNR and RoadLINK tuner phase noise. ×=measurement. $P(|\Delta f_{c,residual}| > \text{accuracy}) = 0.001$

CFO estimation accuracy using auto-correlation depends on the channel between master and slave, the SNR of the received sequence, the phase noise added by the master and by the slave, and the length of the synchronization sequence. As channel we assume Channel D [9]. We can conclude from simulations, which take phase noise of the modem into account, that for SNR > 30 dB the impact of additive noise on the estimation performance is insignificant in respect to the phase noise. We assume that the SNR is at least 30 dB, which is a reasonable assumption based on RF measurements that were conducted with a passenger car and coach bus. In Figure 6 the simulated performance of the CFO estimator is shown vs. the synchronization sequence length. An accuracy of 0.01 PPM requires a synchronization sequence of at least 120 µs.

As synchronization sequence we use 80 repeating blocks of 16 samples (IEEE 802.11 STS) which result in a total length of 128 µs. In order to overcome frequency offset estimation ambiguities in the auto-correlation (AC) results, we estimate the CFO in three stages to obtain a CFO estimator with a range $> \pm 40$ PPM and an accuracy of ± 0.01 PPM as shown in Table 2.

Table 2: CFO estimator stages (with corresponding graphic)

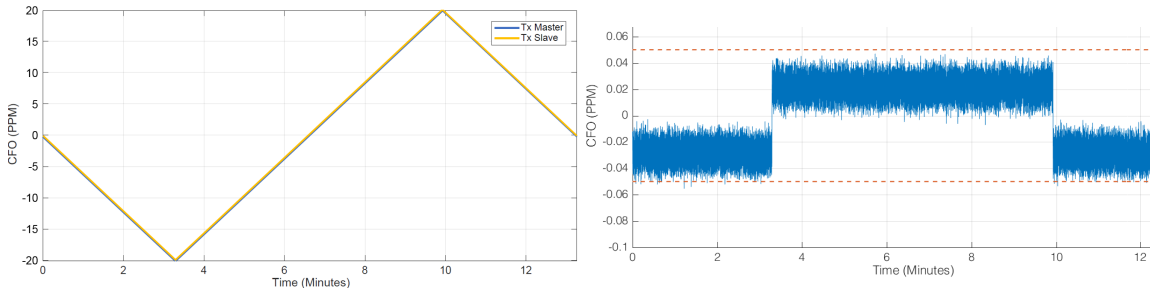
Stage	AC blocks	AC start	AC lag	Range	Accuracy
Coarse	2×1	3	1	± 53 PPM	± 2 PPM
Medium	2×1	3	6	± 8.8 PPM	± 0.33 PPM
Fine	2×38	3	1	± 1.4 PPM	± 0.01 PPM

The PHY scrambler starting seed and media access control (MAC) sequence number counter also need to be synchronized to ensure that both transmitted diversity frames are identical. We do this by including these values in the payload of the synchronization frame, so that the slave can update its values accordingly.

A challenge that still needs to be overcome in a distributed architecture is the problem of synchronizing MAC layer mechanisms like request/clear-to-send, collision avoidance and retransmissions. Furthermore, the synchronization and acknowledgment frames are not compliant to the IEEE 802.11p standard and 5.9 GHz band regulations. Instead they could be exchanged in an ISM band. Another solution to this challenge would be to synchronize the slave on out-going IEEE 802.11p compliant frames of the master, e.g. by correlating these received frames with an equivalent frame generated by the slave itself (the slave has got the payload via the IVN).

4 Results

In this section, performance results of the proposed solution are shown. In Figure 7a it is shown that, when the oscillator frequency of the master is swept between -20 and 20 PPM with 0.01 PPM/s, the slave follows the master frequency over the whole range. In Figure 7b it can be seen that the CFO between master and slave remains within $\pm$ 0.05 PPM.



(a) CFO between master/slave and reference

(b) Residual CFO between master and slave

Figure 7: CFO over time with 0.01 PPM/s drift between -20 and 20 PPM

We compare the performance of our distributed transmit diversity solution with a conventional non-distributed transmit diversity solution by transmitting 5000 packets of 1000 bytes to an IEEE 802.11p receiver that employs channel tracking with 16 µs latency. In Figure 8 the resulting packet error rate (PER) is shown for several averaged

received signal strengths in a flat fading channel for MCS2. Table 3 shows the performance loss of our distributed solution in comparison to the non-distributed solution for each MCS at a PER of 10% in an AWGN and flat fading channel.

MCS	AWGN	Flat fading
0	0.0 dB	0.0 dB
1	0.0 dB	0.0 dB
2	0.1 dB	0.0 dB
3	0.1 dB	0.3 dB
4	0.3 dB	0.3 dB
5	0.2 dB	0.3 dB
6	0.1 dB	0.1 dB
7	-	-0.2 dB

Table 3: Performance loss: distributed vs. non-distributed

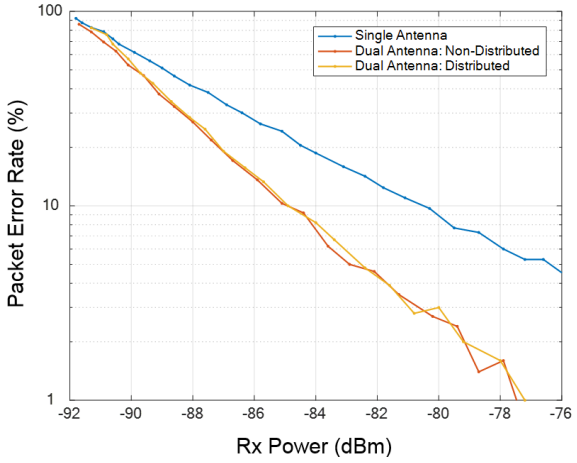


Figure 8: PER MCS2 in flat fading channel

5 Conclusion

Distributed transmit diversity for IEEE 802.11p is feasible with negligible loss of performance by wireless synchronization without the need for additional hardware or a wired interface. We have developed and implemented a wireless synchronization solution on an IEEE 802.11p platform by only reusing available hard- and software resources. With a 340 µs frame exchange, the CFO and TO between the two distributed transceivers is reduced to respectively ± 0.01 PPM and ± 35 ns. By synchronizing every 400 ms we were able to keep the CFO and TO below respectively 0.05 PPM and ± 50 ns when the oscillators have a mutual drift of 0.1 PPM/s. By simulations we have determined that remaining within these limits should not result in a performance degradation in IEEE 802.11p receivers that employ channel tracking with a latency of at most 16 µs. Performance measurement results confirm this and show that our distributed diversity solution can approach the performance of a conventional non-distributed solution with performance penalties of only 0.0-0.3 dB for all MCSs. The IEEE 802.11p standard does not require receivers to employ channel tracking, it is needed to do so in order to perform well in environments with high mobility. Therefore, interoperability could be an issue. Furthermore, the implemented solution is not standard, nor regulations compliant, but we have proposed potential solutions to tackle this challenge.

References

[1] A. Filippi *et al.*, "Why 802.11p beats LTE and 5G for V2x," eeNews Automotive, 2016.

[2] D. K. Borah, G. Moreno-Crespo, and S. Nammi, "Distributed Alamouti Transmit Diversity Technique for Co-Operative Communications," in *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, April 2007, pp. 2210–2214.

[3] M. Klaassen, A. Koppelaar, and P. Alexander, "A Scalable Approach for Low to High End Application of ITS Car2Car Communication," in *19th ITS World Congress*, October 2012.

[4] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Std. 802.11-2012, March 2012.

[5] A. Dammann and S. Kaiser, "Standard conformable antenna diversity techniques for OFDM and its application to the DVB-T system," in *Global Telecommunications Conference, 2001. GLOBECOM '01. IEEE*, vol. 5, 2001, pp. 3100–3105 vol.5.

[6] A. Brakemeier, "ITS-G5 Channel Models for Tx-Diversity," Daimler, Vehicle-Vehicle and Vehicle-Infrastructure Communications, July 2016.

[7] P. Alexander, D. Haley, and A. Grant, "Cooperative intelligent transport systems: 5.9-ghz field trials," *Proceedings of the IEEE*, vol. 99, no. 7, pp. 1213–1235, July 2011.

[8] R. van Nee *et al.*, "The 802.11n MIMO-OFDM Standard for Wireless LAN and Beyond," in *Wireless Personal Communications*, June 2006, pp. 445–453.

[9] V. Erceg *et al.*, *TGn Channel Models*, IEEE Std. 802.11-03/940r4, 2004.

[10] M.-H. Cheng and C.-C. Chou, "Maximum-likelihood estimation of frequency and time offsets in ofdm systems with multiple sets of identical data," *IEEE Transactions on Signal Processing*, vol. 54, no. 7, pp. 2848–2852, July 2006.

A Prototype of Finger-vein Phantom

P. Normakristagaluh[1,2] L.J. Spreeuwers[1] R.N.J. Veldhuis[1]
[1]Data management and Biometrics (DMB) Group
Faculty of EEMCS, University of Twente, Netherlands
[2]Indonesian Institute of Sciences, Indonesia
[1](p.normakristagaluh; l.j.spreeuwers; r.n.j.veldhuis)@utwente.nl
[2]pesi001@lipi.go.id

Abstract

This paper describes the developing a prototype of finger phantom to get modeling on a physical vein pattern. This physical model could help us to get a better understanding of the image formation, since the imaging process seems to be the cause of the poor image quality. The phantom of finger-vein images have been captured using a custom designed capturing device[1]. The set of material considered in this paper consists of 3D printing, soap and titanium dioxide (TiO_2) powder. The prototype of fake bone and tissues within the phantom have already been made from those materials. The various of initial phantoms with and without bone are addressed in this paper as well.

1 Introduction

Finger-vein recognition as a promising biometric technique has drawn increasing attention from the biometrics community in recent years. Compared with other biometric traits, e.g., fingerprint, face or iris, finger-vein characteristics cannot easily be copied, leave no traces around, and are very convenient to use. In anatomy, the vascular pattern is a network structure of veins inside a finger that cannot be damaged easily unless some vein suffers rupture [2].

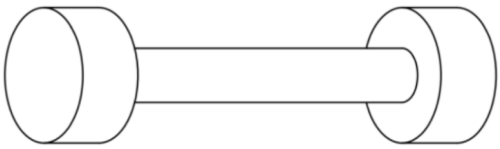
Obtaining a reliable vascular pattern in practice is quite difficult since the images are blurred and have different intensity areas (high-low contrast); for example, figure 2b shows an image of the veins of the middle finger. In fact, strong scattering of light in biological tissue during imaging is the main cause of contrast deterioration in finger-vein images [2]. Until now, knowledge of the image formation process has not really been used in finger vein recognition.

Since the imaging process seems to be the cause of the poor image quality, developing a physical model could help us to get a better understanding of the image formation. We expect that using this model will result in a new robust feature extraction method leading to better recognition performance, which will have an impact on the reliability and usability of finger veins as personal authentication in a real application. A further advantage of a physical model is being able to generate realistic images with a ground truth for the vein patterns, which in turn may lead to better feature extraction. Moreover, if we can improve modeling, we can work on better pre-processing and a better algorithm to visualize finger veins.

In this work, the experiment has three phase of developing the phantom. Developing a fake bone and tissues have already been done as the first two stages. This paper illustrates the preliminary results of those phases with the initial methods and does not describe in detail for the third phase that is creating a fake vein.

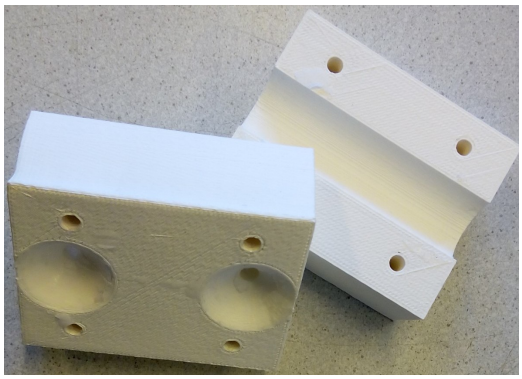
(a) Transparent and white soap material

(b) Titanium dioxide (TiO_2) powder

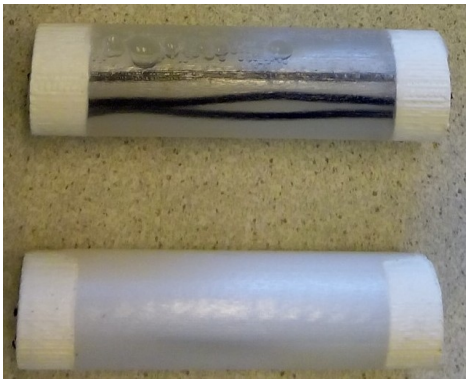


(c) Sketsa of the skeleton fake finger

(d) Skeleton of fake finger



(e) Mold



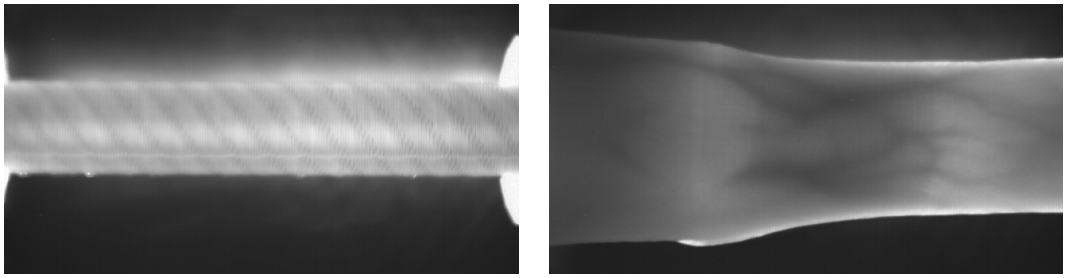
(f) A sample of fake finger

Figure 1: The basis of phantom material.

2 Methods

Tissues as in the case for human skin are heterogeneous and are often composed of different structures having different optical properties. It describes the penetration, absorption, scattering, and remittance of light at different wavelengths. In vivo, veins and arteries have a different sizes and at different depths [3]. Basically, human perception is limited to the visible (VIS) spectral range that is defined by the luminous efficiency functions ranging between wavelengths of $\lambda = 380$ nm and $\lambda = 780$ nm. However, near infra red (NIR) radiation in the tissue and the absorption by haemoglobin in the blood vessels is the best between $\lambda = 850$ and 900 nm. Therefore, in the NIR images, the observable size of the veins most probably is strongly increased [4].

In the same way, the tissue phantom should have optical properties close to the living tissue of human body and tissue components (e.g., epidermis and dermis) [5]. Besides that, phantoms also consist of a scattering medium, an absorbing medium, a diluent, and in some cases, fluorophores. For example, absorbing media include black India ink or/and some biological dyes, such as trypan blue and photofrin. The most common materials providing scattering were presented as: (1) lipid-based emulsions,



(a) The fake bone. (b) The left middle finger.

Figure 2: The Illumination process pass through the fake bone, and the left middle finger as image reference.

(2) titanium or aluminum oxide powders, and (3) polymer microspheres [6]. In fact, titanium oxide nanoparticles are excellent scatters and the most common choice for scattering in science and engineering [5]. In this experiment, titanium dioxide (TiO_2) powder has been used as scatterers, and soap material as an absorber in constructing phantoms. Apart from the finger, bone has been modelled by 3D printing as a base material. All materials supporting in this experiment are presented in figure 1.

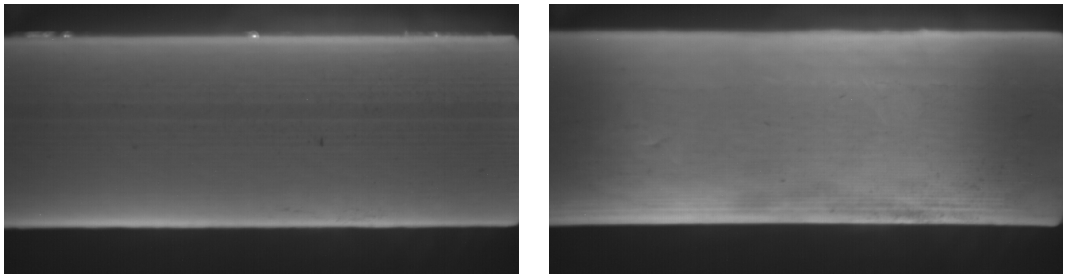
Fabrication of the phantom samples with and without a fake bone was provided by the following steps: melting the soap material in the pan at 90^0C, advanced manual mixing of soap with TiO_2, further mixing of TiO_2/soap solution in room temperature during 5 minutes, pouring of prepared solution into mold finger form and drying the sample in the same temperature within 1 hrs. This phantom, which is mimicking the tissue with and without a fake bone, is the initial phase to get a similar properties with the real tissue of the finger.

3 Preliminary Results and Discussion

The first artificial finger was made from a mixture of white and transparent soap materials, which have a bit similar properties to the biological tissues of the finger. Nevertheless, the illuminations penetrate through this fake tissue within the fake finger. Apart from a fake finger, we have also created a fake bone and a fake vein from thin wire. Figure 1f shows a sample of the first fake finger.

Three issues were encountered during making the phantom. The first was related to uniform the illumination. This can be dealt with by setting the LED on capturing device. Figure 2 shows the images, which have uniform the illumination, captured by the finger scanner device. The second stems from the physiology illumination of bone structure. Theoretically, the real bone should have properties, which is blocking the illumination through the finger. However, the results do not support the hypothesis: from figure 2a it can be seen that the light penetrate through the fake bone. This occurs because the bone material does not have a solid structure. A solution to this problem might be finding the massive material which can block the lighting.

The last issue emerged from scattering process within the fake tissue. These processes are presented in figure 3a and 3b. The images were generated by the same illumination as in figure 2b. It can be seen from figure 3b that the fake tissue without the fake bone inside have fairly similar properties with the joint area in the finger, which is the brightest area in figure 2b. Therefore, this material composition would be base of the fake finger. If all issues solved, the research would continue to obtain the fake of vein structure inside the fake finger.



(a) Fake bone inside (b) No fake bone inside

Figure 3: The capturing fake finger images ware made up of 50 gram soap transparent and 0.5 gram TiO_2.

4 Conclusion

A new method for finger-vein pattern imaging, based on the physical model has been presented. The modeling finger with phantom has three steps on this research. The first and the second step have already been done. These steps are creating a fake bone and tissue. Even though the fake bone in the figure 2a have little bit similar properties with the real bone, but the fake tissues almost have the same optical structure as the biological tissue in the finger. However, this is still preliminary result of the research. The more accurate composition of materials should be developed to obtain the same properties as the real finger. The next step is developing physical model of vein to get the entire structure of finger phantom.

References

[1] B. T. Ton and R. N. Veldhuis, "A high quality finger vascular pattern dataset collected using a custom designed capturing device," in *Biometrics (ICB), 2013 International Conference on*. IEEE, 2013, pp. 1–5.

[2] J. Yang and Y. Shi, "Towards finger-vein image restoration and enhancement for finger-vein recognition," *Information Sciences*, vol. 268, pp. 33 – 52, 2014, new Sensing and Processing Technologies for Hand-based Biometrics Authentication. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025513007342

[3] N. J. Cuper, J. H. Klaessens, J. E. Jaspers, R. de Roode, H. J. Noordmans, J. C. de Graaff, and R. M. Verdaasdonk, "The use of near-infrared light for safe and effective visualization of subsurface blood vessels to facilitate blood withdrawal in children," *Medical Engineering and Physics*, vol. 35, no. 4, pp. 433–440, 2013.

[4] K. Mangold, J. A. Shaw, and M. Vollmer, "The physics of near-infrared photography," *European Journal of physics*, vol. 34, no. 6, p. S51, 2013.

[5] V. Tuchin, A. Bashkatov, E. Genina, V. Kochubey, V. Lychagov, S. Portnov, N. Trunina, D. Miller, S. Cho, H. Oh *et al.*, "Finger tissue model and blood perfused skin tissue phantom," in *Dynamics and Fluctuations in Biomedical Photonics VIII*, vol. 7898. International Society for Optics and Photonics, 2011, p. 78980Z.

[6] V. V. Tuchin and V. Tuchin, "Tissue optics: light scattering methods and instruments for medical diagnosis," 2007.

Region of interest segmentation of VLE data using CNN and weighted groundtruth

Joost van der Putten[A], Fons van der Sommen[A], Maarten Struyvenberg[B], Jeroen de Groof[B]
Wouter Curvers[C], Erik Schoon[C], Jaques J.G.H.M. Bergman[B], Peter H.N. de With[A]
[A] Eindhoven University of Technology, 5612 AP, Eindhoven, the Netherlands;

[B] Academic medical center, 1105 AZ, Amsterdam, the Netherlands;
[C] Catharina Hospital, 5623 EJ, Eindhoven, the Netherlands
j.a.v.d.putten@tue.nl

Abstract

Volumetric Laser Endomicroscopy (VLE) is a promising balloon based imaging technique for detecting early neoplasia in Barrett's Esophagus. Especially Computer Aided Detection (CAD) techniques show great promise compared to doctors, who cannot reliably find disease patterns in the VLE signal. However, the relevant tissue has to be segmented in order for these systems to function properly. At present, tissue segmentation has to be done manually and is therefore not scalable for full VLE scans of 1,200 × 4,096 × 2,048 pixels. Furthermore, the current CAD methods cannot use the VLE scans to their full potential as only a small section is selected while an automated system can delineate the entire image. This paper explores the possibility of automatically segmenting relevant tissue for VLE scans using a convolutional neural network. The contribution of this work is twofold. First, this is the first tissue segmentation algorithm for VLE scans. Second, we introduce a weighted ground truth that exploits the signal to noise ratio characteristics of the data. The results show that our approach achieves excellent results that are comparable to human annotations, using a weighted groundtruth slightly increases the segmentation accuracy of the algorithm as well.

1 Introduction

Esophageal adenocarcinoma (EA) is a form of cancer whose incidence is rising dramatically in the Western world. Patients with Barrett's Esophagus (BE) have an increased risk for developing dysplasia and eventually EA. Hence, it is crucial that dysplasia associated with BE is detected at an early stage. Currently, BE patients are monitored through periodic endoscopic surveillance and biopsies. However, the current methods have some drawbacks such as diagnostic uncertainty of white-light endoscopy, sampling error, and ambiguous samples in histology. Volumetric Laser Endomicroscopy (VLE) is a novel technique that has the potential to significantly contribute to the early detection of dysplsia. A recent investigation into CAD for VLE analysis showed promising results, with an Area Under Curve (AUC) between 90%-93% [1]. However, in order to use the proposed methods, the tissue under examination has to be extracted, which is currently done by manual cropping. This technique is limited, as not all valuable information is extracted from the images. Additionally, this process is not scalable, since full VLE scans consist of 1,200 slices of 4k × 2k pixels per patient.

In this research we propose the first tissue segmentation algorithm for VLE images, using U-net [2] with an adapted domain-specific loss function.

(a) (b) (c) (d)

Figure 1: (a) VLE image. (b) Basic ground truth. (c) Weighted ground truth. (d) Prediction output of the CNN, using the weighted ground truth to train the U-net.

2 Methods

In order to segment the VLE images, we implement an end-to-end learning strategy. First, the ground truth needs to be determined prior to using the images to train a CNN. The various experiments are then evaluated against both general and domain-specific metrics.

2.1 Ground truth labeling

Two different sets of ground truth data were used for the training and evaluation of the algorithm. The first ground truth was manually delineated by a CAD expert with more than three years experience in VLE research, which will be further referenced as the basic ground truth. An example of a VLE image and its ground truth are shown in Figure 1a and Figure 1b, respectively. The strength of the VLE signal weakens considerably over the depth, thereby decreasing the signal-to-noise ratio in the lower parts of the images. Additionally, research has indicated that incident cancer is most discriminative in the top layers of tissue [3]. Hence, the top layers are more important than the lower layers. For this reason, a weighted ground truth was implemented as well. The uppermost border is copied from the first basic ground truth masks and then an additional region below that border is added to the mask, reducing the mask weight for lower layers. An example is shown in Figure 1c.

2.2 Convolutional neural network (CNN)

A two-class CNN similar to the one described by Ronneberger *et al.* [2] was used to segment the tissue of the VLE scans. The original snapshots have a resolution of 1,342 $\times$ 1,024 pixels, which were resized to 256 $\times$ 256 pixels for computational efficiency. The final layer has a sigmoid activation function, so that the output represents the probability whether a pixel belongs to the segmented tissue or not. For the loss function, the Dice Similarity Coefficient (DSC) was used. Generally, the DSC is calculated using two binary masks. However, a threshold is necessary in order to binarize the masks, making the DSC non-differentiable, which is required by the back-propagation algorithm. These conflicting demands were solved by implementing the DSC with parameter y as the ground truth, $\hat{y}$ being the prediction and $y, \hat{y} \in [0, 1]$, as follows:

$$DSC = \frac{2(y \cdot \hat{y})}{(y + \hat{y})}.$$

2.3 Evaluation

The applied data set for this research consisted of 137 histopathologically matched VLE snapshots. Fourfold cross-validation was performed to calculate the metrics over the entire data set. Prior to calculating the evaluation metrics, the weighted ground truth and the predicted segmentation were first binarized with the Otsu threshold to facilitate fair comparisons. The predicted tissue segmentation was compared to the ground truth by calculating the binary DSC and a custom metric, called balloon distance.

The balloon distance is calculated by first taking the difference between the binarized weighted ground truth and the prediction. Since a correct upper boundary is critical for further classification tasks, only the wrongly classified pixels in the top region are considered. We have defined the balloon distance as follows:

$$\text{Balloon distance} = 1 - min(N_z(\text{difference mask})/1000, 1),$$

where N_z denotes the sum of the non-zero elements in the difference mask. Since the balloon in the resized ground truth has an approximate height of 3 or 4 pixels and spans the entire width of 256 pixels, the image was rescaled with a value of 1000 to normalize the results. This yields in a metric where unity indicates that the top part of prediction aligns perfectly with the top part of the ground truth, and zero indicates a total mismatch between the ground truth and the prediction.

3 Experimental Results

The results of our experiments are shown in Table 1, where we employed 10,000 epochs of training for all experiments. Both the DSC and the Balloon Distance (BD) were evaluated, using the predictions obtained from the model generated from the corresponding ground truth type. An example of a prediction made by the algorithm is shown in Figure 1d. For this example, the network was trained with the weighted ground truth DSC as a loss function.

	Basic model		Weighted model	
	DSC	BD	DSC	BD
Assessor 1	0.95	0.88	0.97	0.88
Assessor 2	0.95	0.83	0.97	0.83
Assessor 3	0.95	0.85	0.97	0.85

Table 1: DSC and balloon distance (BD) comparison of the different assessors and the weighted ground truth and basic ground truth predictions, obtained with their respective model.

4 Conclusions

Training a network with a weighted ground truth increases the DSC by approximately 2%, resulting in tissue segmentations that are nearly identical to the human annotations. However, using a weighted groundtruth has no significant effect on the ability of the model to precisely delineate the topmost layer (BD metric) of the relevant VLE tissue. This is especially true for the BD metric. In this work, we have proposed the first tissue segmentation algorithm for VLE scans, as well as a domain-specific loss function that exploits the signal-to-noise characteristics of VLE scans. Interestingly, our type

of data exploitation and characteristics learning is of generic nature and can therefore also be implemented for other similar modalities, such as ultrasound and other optical coherence tomography applications.

Acknowledgments

We would like to thank the Academic Medical Center in Amsterdam and the Catherina Hospital in Eindhoven, who provided the medical scans used in this research, as well as provide crucial insights into the data. We thank NVIDIA for donating a Titan-XP GPU.

References

[1] F. van der Sommen et al, Predictive features for early cancer detection in Barrett's esophagus using volumetric laser endomicroscopy, Comput Med Imaging Graph, 2018, [in press].
[2] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, pp. 18, 2015.
[3] A. F. Swager et al., Computer-aided detection of early Barretts neoplasia using volumetric laser endomicroscopy, Gastrointest. Endosc., vol. 86, no. 5, pp. 839846, 2017.

Calculation of the Mean Strain of Non-uniform Strain Fields Using Conventional FBG Sensors

Aydin Rajabzadeh[a,b], Roger M. Groves[b], Richard C. Hendriks[a], and Richard Heusdens[a]

[a] Circuits and Systems group of Delft University of Technology, Delft, The Netherlands
[b] Structural Integrity and Composites group of Delft University of Technology, Delft, The Netherlands
{a.rajabzadehdizaji, r.m.groves, r.c.hendriks, r.heusdens}@tudelft.nl

Abstract

Fibre Bragg grating sensors are spatially modulated patterns of refractive index changes in the core of optical fibres that reflect a narrow wavelength spectrum of the input light. The main application of FBG sensors is in the field of distributed point strain measurement. With the possibility of multiplexing several sensors on a single optical fibre and their immunity to electromagnetic interferences, FBG sensors are an ideal candidate for applications in adverse environments such as the oil and gas industries or aviation. One of the most interesting properties of these sensors is the presumed linear relationship between the strain and the peak wavelength shift of the FBG reflected spectra. However, subjecting sensors to a non-uniform stress field will in general result in a strain estimation error when using this linear relationship. Despite the fact that in many practical situations the strain distribution is non-uniform, existing algorithms for estimating the mean strain value are still based on the shift of the peak wavelength. When the sensor is subject to non-uniform strain fields, however, each segment of the sensor will experience a different strain, resulting in different peak wavelength shifts along the length of the sensor. In order to analyse the reflected spectra under such non-uniform stress fields, we will present a model that describes the interaction of the forward and backward electric wave propagation between consecutive segments, which is an approximation of the widely used transfer matrix model (TMM). This approximated transfer matrix model (ATMM) enables us to accurately find the average wavelength shift.

In the ATMM we assume that the length of the sensor is divided into several piece-wise uniform segments whose lengths are sufficiently small. Under this condition, we show that it is possible to approximate the side-lobes of the reflected spectra with a closed form expression. This closed-form expression can be exploited to devise an algorithm that is a maximum likelihood estimator of the mean strain of "smooth" non-uniform strain fields. Taking into account the practical considerations associated with real FBG sensors, including the production defects and birefringence effects, the proposed algorithm is summarised as

1. Align the main peaks of the reflected spectra of the FBG sensor measured with and without applying a (non-uniform) strain, thereby defining $\Delta\lambda_B$, the wavelength shift of the main peak.

2. Maximise the cross-correlation of the side lobes of both measurements over a small interval around $\tilde{\lambda} = \lambda_B + \Delta\lambda_B$, resulting in an additional phase shift $\delta\lambda_B$.

3. Calulate the required phase shift $\bar{\lambda}_B - \lambda_B = \Delta\lambda_B + \delta\lambda_B$.

4. Calculate the mean strain based on the new compensated wavelength shift with $\bar{s} = (\bar{\lambda}_B - \lambda_B)/k_s$, where $\bar{s}$ is the mean strain and k_s is a constant that is dependent on the physical properties of the sensor.

In other words, we propose an alternative two step algorithm where we first measure the Bragg wavelength shift as is done in traditional strain estimation algorithms, and then refine the estimate by cross-correlating the side lobes of both spectra over a small range around the shifted Bragg wavelength. We validated this proposed algorithm using both computer simulations and experimental FBG measurements and showed that the newly proposed algorithm clearly outperforms state-of-the-art strain estimation algorithms by compensating for mean strain errors of around $30\mu\varepsilon$. Figure 1 demonstrates the result of the above mentioned steps

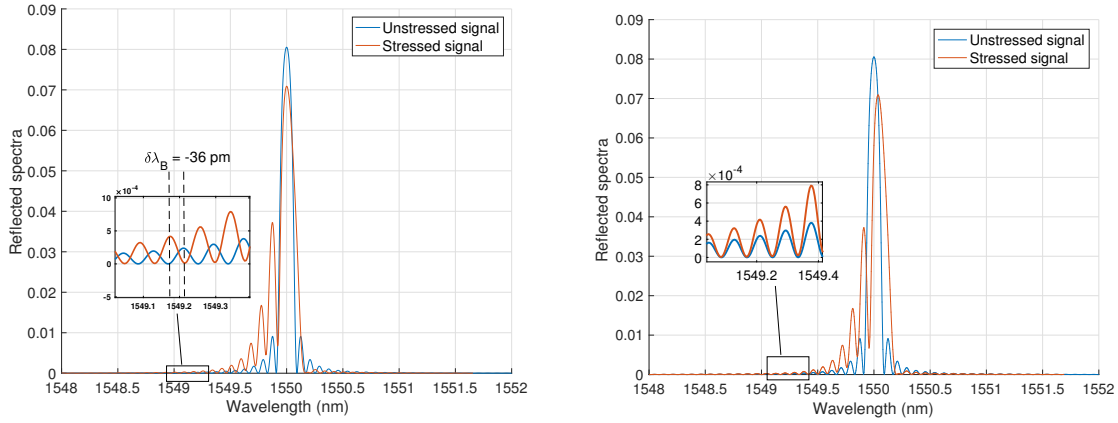


Figure 1: The FBG reflected spectra of the unstrained sensor (blue) and the strained sensor (red) where (a) shows both spectra aligned with respect to their Bragg wavelength (the stressed signal is already shifted for $\Delta\lambda_B = 344$pm) and (b) with respect to maximising the cross-correlation of the side peaks. The required phase shift in this example is $\bar{\lambda}_B - \lambda_B = 344 - 36 = 308$pm.

Social diversity for reducing the impact of information cascades on social learning

Fernando Rosas[1,2,3], Kwang-Cheng Chen[4] and Deniz Gündüz[2]
[1] Department of Mathematics, Imperial College London, UK
[2] Department of Electrical and Electronic Engineering, Imperial College London, UK
[3] Centre of Complexity Science, Imperial College London, UK
[4] Department of Electrical Engineering, University of South Florida, USA
f.rosas@ic.ac.uk kwangcheng@usf.edu d.gunduz@ic.ac.uk

Abstract

Collective behavior in online social media and networks is known to be capable of generating non-intuitive dynamics associated with crowd wisdom and herd behaviour. Even though these topics have been well-studied in social science, the explosive growth of Internet computing and e-commerce makes urgent to understand their effects within the digital society. In this work we explore how the stochasticity introduced by social diversity can help agents involved in a inference process to improve their collective performance. Our results show how social diversity can reduce the undesirable effects of information cascades, in which rational agents choose to ignore personal knowledge in order to follow a predominant social behaviour. Situations where social diversity is never desirable are also distinguished, and consequences of these findings for engineering and social scenarios are discussed.

1 Introduction

The high interconnectedness enabled by communication technologies and online media is progressively increasing the complexity of our aggregated social behaviour [1]. In fact, these complex dynamics were dramatically illustrated by the failure of our prediction tools in the forecast of recent political events, including the Brexit referendum and the latest US presidential election. A key open challenge is to clarify how the large amount of information that is constantly exchanged among individuals affects their decisions.

Fascinating dynamics take place when social agents engage in sequencial decision-making. For example, most people nowadays use the Internet to check other people's recommendations prior to make decisions, which enable more informed decisions thanks to the inclusion of evidence from previous experiences. Subsequent decisions are, however, heavily influenced by earlier agents, allowing misinformation or fake news to be reinforced and spread across the social network. These non-trivial *social learning* dynamics are known to play a critical role in a number of key social phenomena, e.g., in the adoption or rejection of new technology, and in the formation of political opinions [2,3]. Moreover, social learning also plays a key role in the context of e-commerce and digital society, e.g., in recommendation systems of online stores where users access opinions of previous customers while choosing their products [4, 5]. This is also the case in the emergence of viral media contents in various Internet portals, which are based on sequential actions of like or dislike.

A deep understanding of social learning dynamics is crucial for enabling robust platform design against fake news and data falsification, which is an urgent need in our modern networked society. As a matter of fact, digital misinformation was listed by the World Economic Forum (WEF) as one of the main threats to our modern society [6].

Social learning have been thoughtfully studied since the 90's by researchers from economics and social sciences [7–9] (for modern reviews see [2, 3]). These studies have

shown that social learning is driven by two competing mechanisms. In one hand, the well-known *crowd wisdom* improve the desicion-making capabilities of agents within large networks, as more information becomes available to latter agents in the desicion sequence. The accumulation of social experience can, on the other hand, overload agents and generate *information cascades*, which pushes them to ignore their private knowledge and to adopt a predominant social behaviour. Interestingly, it has been shown that the combination of these two mechanisms can serve to provide network resilience against data falsification attacks [10, 11], pointing out promising possibilities for the design of resilient social learning platforms.

Motivated by the benefits that diversity can provide in biological and social systems [12, 13], in this work we study how social diversity affects the learning rate in a social learning scenario. For this, we consider a network of rational agents that have diverse preferences and prior information, having some similarities to the works reported in [14, 15]. Using a communication theoretical interpretation of this scenario, we show that social diversity is equivalent to additive noise in a communication channel —which one would expect to be detrimental for the learning process. Surprisingly, our findings show that social diversity can help to avoid information cascades, introducing important improvements in the asymptotic learning performance.

The rest of this article is structured as follows. Sections 2 introduces the considered social learning scenario, and develops our definition of social diversity. Section 3 defines information cascades, and characterize theoretically their behaviour with respect to social diversity. Section 4 presents numerical evaluations that verify the theoretical results, and finally Section 5 summarizes our main conclusions.

Notation: uppercase letters X are used to denote random variables and lowercase x realizations of them, while boldface letters $\boldsymbol{X}$ and $\boldsymbol{x}$ represent vectors. Also, $\mathbb{P}_w\{X = x|Y = y\} \coloneqq \mathbb{P}\{X = x|Y = y, W = w\}$ is used as a shorthand notation.

2 Social learning model

2.1 Preliminaries and basic assumptions

Let us consider a social network composed by N agents, who are engaged in a decision-making process. In this process each agent need to make a decision between two options*, which could correspond to a choice between two restaurants, two brands, or two political parties. It is assumed that decisions occur sequentially, and are labeled according to the order in which they take place.

The decision of the n-th agent, denoted as $X_n \in \{0, 1\}$, is based on two sources of information (see Figure 1): a *private signal* $S_n \in \mathcal{S}$, which is a continuous or discrete random variable that represents personal information that the n-th agent possesses, and *social information* given by the decisions of the previous agents, denoted by $\boldsymbol{X^{n-1}} \coloneqq (X_1, \ldots, X_{n-1}) \in \{0, 1\}^{n-1}$.

All the agents are assumed to have equivalent observation capabilities, and therefore the private signals S_n are identically distributed. These signals are affected by environment conditions, which for simplicity are represented by a binary variable W. For the sake of tractability, we follow the existent literature in assuming that the private signals S_n are conditionally independent given W, leaving other cases for future work. The corresponding conditional probability distributions of S_n given the event $\{W = w\}$ are denoted by μ_w. We further assume that no realization of S_n is capable of completely determining W, which is equivalent to the measure theoretic notion of absolute continuity between μ_0 and μ_1 [16]. As a consequence of this assumption, the

*Although generalizations for more than two options are possible, we focus in the case of binary decisions for simplifying the presentation.

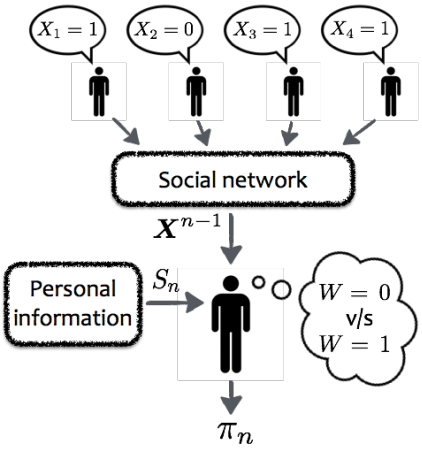


Figure 1: A social learning scenario, where an agent needs to make a decision (π_n) based on personal information coming from a private signal (S_n) and social information ($\boldsymbol{X}^{n-1}$) coming from a social network.

log-likelihood ratio of these two distributionsμ_1 and μ_0 is well-defined and given by the logarithm of the corresponding Radon-Nikodym derivative $\Lambda_S(s) = \log \frac{d\mu_1}{d\mu_0}(s)^\dagger$.

A *strategy* is a rule for generating a decision X_n based on $S_n = s$ and $\boldsymbol{X}^{n-1}$, i.e. a collection of deterministic or random functions $\pi_n : \mathcal{S} \times \{0,1\}^{n-1} \to \{0,1\}$ such that $X_n = \pi_n(S_n, \boldsymbol{X}^{n-1})$.

2.2 Bayesian strategy, agents' preferences and prior information

Let us assume that the preferences of the n-th agent are encoded in an utility function $u_n(x, w)$, which determines the payoff that the agent receives when making the decision $X_n = x$ under the condition $\{W = w\}$. We consider rational agents that follow a *Bayesian strategy*, which seeks to maximize their average payoff given by $\mathbb{E}\left\{u(\pi_n(S_n, \boldsymbol{X}^{n-1}), W)\right\}$, with $\mathbb{E}\{\cdot\}$ being the expected value operator. It has been shown that the Bayesian strategy for the n-th agent can be expressed succinctly as [4]

$$\frac{\mathbb{P}\left\{W = 1 | S_n, \boldsymbol{X}^{n-1}\right\}}{\mathbb{P}\left\{W = 0 | S_n, \boldsymbol{X}^{n-1}\right\}} \underset{X_n=1}{\overset{X_n=0}{\lessgtr}} e^{\nu_n} \quad , \tag{1}$$

where $\nu_n = \log \frac{u_n(0,0) - u_n(0,1)}{u_n(1,1) - u_n(1,0)}$ reflects the effect of the cost function. For considering agents with diverse preferences, we assume that ν_i are independent and identically distributed (i.i.d.) random variables.

Let us further consider the case where the agents have no absolute knowledge about the prior distribution of W. Note that because W is binary, its distribution is completely determined by the value of $\mathbb{P}\{W = 1\}$. Following the framework of Bayesian inference [17], let us consider $\theta_n \in [0, 1]$ to be a collection of i.i.d. random variables following a distribution $f_\theta(\theta)$ that reflects the state of knowledge of the agents about $\mathbb{P}\{W = 1\}$. In particular, if the agent has complete knowledge then $f_\theta(\theta)$ is a delta

†When S_n takes a discrete number of values then $\frac{d\mu_1}{d\mu_0}(s) = \frac{\mathbb{P}\{S_n = s | W=1\}}{\mathbb{P}\{S_n = s | W=0\}}$, while if S_n is a continuous random variable with conditional p.d.f. $p(s|w)$ then $\frac{d\mu_1}{d\mu_0}(s) = \frac{p(s|w=1)}{p(s|w=0)}$.

centered in the true value of $\mathbb{P}\{W=1\}$ and hence $\theta_n = \mathbb{P}\{W=1\}$ for all n, while if agents has no information then $f_\theta(\theta)$ corresponds to an uniform distribution over $[0,1]$.

Noting that $\boldsymbol{X}^{n-1}$ depends only on $(S_1,\ldots,S_{n-1})$, and therefore is conditionally independent of S_n, a direct application of the Bayes rule on $\mathbb{P}\left\{W=1|S_n,\boldsymbol{X}^{n-1}\right\}$ and $\mathbb{P}\left\{W=0|S_n,\boldsymbol{X}^{n-1}\right\}$ shows that (1) can be re-written as

$$\Lambda_S(S_n) + \Lambda_{\boldsymbol{X}^{n-1}}(\boldsymbol{X}^{n-1}) \underset{X_n=1}{\overset{X_n=0}{\lessgtr}} \nu_n + \log\frac{\theta_n}{1-\theta_n} \ , \tag{2}$$

where $\Lambda_S(S_n)$ and $\Lambda_{\boldsymbol{X}^{n-1}}(\boldsymbol{X}^{n-1})$ are the log-likelihood ratios of S_n and $\boldsymbol{X}^{n-1}$, respectively. Note that an efficient method for computing $\tau_n(\boldsymbol{X}^{n-1})$ has been reported in [10].

2.3 Communication theoretic interpretation

By using an adequate decision labeling, one can consider the event $\{X_n = W\}$ to be more desirable than $\{X_n \neq W\}$, or equivalently, that $u_n(1,1) \geq u_n(1,0)$ and $u_n(0,0) \geq u_n(0,1)$. The Bayesian strategy is, hence, to choose X_n as similar to W as possible using the information provided by S_n and $\boldsymbol{X}^{n-1}$. Therefore, the decisions $\pi_n(S_n,\boldsymbol{X}^{n-1}) = X_n$ can be considered to be noisy estimations of W.

To further explore this perspective, let us re-formulate (2) as

$$\Lambda_S(S_n) + \xi_n \underset{X_n=1}{\overset{X_n=0}{\lessgtr}} \tau_n(\boldsymbol{X}^{n-1}) \ , \tag{3}$$

where $\xi_n := \log(1-\theta_n)/\theta_n - \nu_n$ and $\tau_n(\boldsymbol{X}^{n-1}) := -\Lambda_{\boldsymbol{X}^{n-1}}(\boldsymbol{X}^{n-1})$. The above can be understood as a classic signal decoder within communication theory [4, Section IV], where $\Lambda_S(S_n)$ is the decision signal and ξ_n is additive noise. Moreover, $\tau_n(\boldsymbol{X}^{n-1}$ is a decision threshold that establishes the decoding rule based on a Vonoroi tessellation that divides $\mathbb{R}$ in two semi-open intervals given by $(-\infty,\tau_n(\boldsymbol{X}^{n-1}))$ and $(\tau_n(\boldsymbol{X}^{n-1}),\infty)$.

3 Avoiding information cascades via noise

3.1 Local and global information cascades

In general, the decision $\pi_n(S_n,\boldsymbol{X}^{n-1})$ is made based in complementary evidence provided by both $\boldsymbol{X}^{n-1}$ and S_n. The n-th agent is said to fall into a *local information cascade* when the information conveyed by S_n is not included in the decision-making process due to a dominant influence of $\boldsymbol{X}^{n-1}$. The term "local" is used to emphasize that this event is related to the data fusion taking place at an individual agent. The notion of local information cascade is formalized in the following definition, which is based on the notion of conditional mutual information [18], denoted as $I(\cdot;\cdot|\cdot)$.

Definition 1. *The social information* $\boldsymbol{x}_c^{n-1} \in \{0,1\}^{n-1}$ *generates a* local information cascade *for the n-th agent if* $I(\pi_n; S_n|\boldsymbol{X}^{n-1} = \boldsymbol{x}_c^{n-1}) = 0$.

The above condition summarizes two possibilities: either $\pi_n(s,\boldsymbol{x}_c^{n-1})$ is constant for all values of $s \in \mathcal{S}$, or there is still variability but this variability is conditionally independent of S_n (e.g. in the case of stochastic strategies —not considered in this work). In both cases, the above definition highlights the fact that the decision π_n contains no unique information[‡] coming from S_n when a local cascade takes place .

[‡]For a rigorous definition of unique information in Markov chains c.f. [19].

Next we define *global information cascades*, which are avalanches of local information cascades that affect all the agents after their ignition.

Definition 2. *The social information vector* $\boldsymbol{x}_c^{n-1} \in \{0,1\}^{n-1}$ *triggers a* global information cascade *if* $I(\pi_m : S_m | \boldsymbol{X}^{n-1} = \boldsymbol{x}_c^{n-1}) = 0$ *holds for all* $m \geq n$.

The relationship between local and global information cascades is explored in the next section (c.f. Proposition 1).

3.2 The effect of social diversity over information cascades

Let us first introduce $F_w(z) = \mathbb{P}_w \{\Lambda_S(S_n) + \xi_n \leq z\}$ as a shorthand notation for the cumulative distribution function of $\Lambda_S(S_n) + \xi_n$ conditioned on the event $\{W = w\}$. Note that, thank to the fact that $\Lambda_S(S_1)$ and ξ_1 are independent random variables, one can compute $F_w(\cdot)$ as the convolution of their density functions.

Lemma 1. *The conditional statistics of* π_n *given* $\boldsymbol{X}^{n-1}$ *are defined by*

$$\mathbb{P}_w \left\{ \pi_n = 0 | \boldsymbol{X}^{n-1} = \boldsymbol{x}^{n-1} \right\} = F_w(\tau_n(\boldsymbol{x}^{n-1})). \tag{4}$$

Proof. A direct calculation shows that

$$\mathbb{P}_w \{\pi_1(S_1) = 0\} = \mathbb{P}_w \{\Lambda_S(S_1) + \xi_1 < 0\} = F_w(0).$$

Following a similar rationale, one can find that

$$\begin{aligned}
\mathbb{P}_w \left\{ \pi_n = 0 | \boldsymbol{X}^{n-1} = \boldsymbol{x}^{n-1} \right\} &= \int_{\mathcal{S}} \mathbb{P}_w \left\{ \pi_n(s, \boldsymbol{x}^{n-1}) = 0 | S_n = s \right\} \mu_w(s) \mathrm{d}s \\
&= \int_{\mathcal{S}} \mathbb{1} \left\{ \pi_n(s, \boldsymbol{x}^{n-1}) = 0 \right\} \mu_w(s) \mathrm{d}s \\
&= \mathbb{P}_w \left\{ \Lambda_S(s) + \xi_n < \tau_n(\boldsymbol{x}^{n-1}) \right\} \\
&= F_w(\tau_n(\boldsymbol{x}^{n-1})) \ .
\end{aligned}$$

Above, the first equality is a consequence of the fact that S_n is conditionally independent of $\boldsymbol{X}^{n-1}$ given $W = w$, while the second equality is a consequence that π_n is a deterministic function of $\boldsymbol{X}^{n-1}$ and S_n, and hence becomes conditionally independent of W. □

Next, using Lemma 1, one can show that τ_n is an effective summary of the information provided by $\boldsymbol{X}^{n-1}$ that is relevant for generating the decision π_n.

Lemma 2. *The variables* $\boldsymbol{X}^{n-1} \to \tau_n \to \pi_n$ *form a Markov Chain, i.e.* τ_n *is a sufficient statistic of* $\boldsymbol{X}^{n-1}$ *for predicting the decision* π_n.

Proof. Using (4), one can find that

$$\mathbb{P}_w \left\{ \pi_n = 0 | \tau_n, \boldsymbol{X}^{n-1} \right\} = F_w(\tau_n) = \mathbb{P}_w \left\{ \pi_n = 0 | \tau_n \right\} \ , \tag{5}$$

and therefore the conditional independency of π_n and $\boldsymbol{X}^{n-1}$ given τ_n is clear. □

We now present a proposition that clarifies the relationship between local and global information cascades. This result extends [4, Theorem 1] to the current scenario.

Proposition 1. *Each local information cascade triggers a global information cascade over the social network.*

Proof. Letus first note that

$$\begin{aligned}\tau_{n+1}(\boldsymbol{X}^n) - \tau_n(\boldsymbol{X}^{n-1}) =& \Lambda_{\boldsymbol{X}^{n-1}}(\boldsymbol{X}^{n-1}) - \Lambda_{\boldsymbol{X}^n}(\boldsymbol{X}^n) \\ =& -\Lambda_{X_n|\boldsymbol{X}^{n-1}}(X_n|\boldsymbol{X}^{n-1}) \ , \end{aligned} \tag{6}$$

where the conditional log-likelihood is given by

$$\Lambda_{X_n|\boldsymbol{X}^{n-1}}(X_n|\boldsymbol{X}^{n-1}) = \log \frac{\mathbb{P}_1\left\{X_n|\boldsymbol{X}^{n-1}\right\}}{\mathbb{P}_0\left\{X_n|\boldsymbol{X}^{n-1}\right\}}.$$

Let us consider $\boldsymbol{x}_{\mathrm{c}}^{n-1} \in \{0,1\}^{n-1}$ such that it produce a local cascade in the n-th node. As Bayesian strategies are deterministic, local information cascades corresponds to the events where π_n is fully determined by $\boldsymbol{X}^{n-1}$, i.e. when the probability of the event $\{\pi_n = 0|\boldsymbol{X}^{n-1} = \boldsymbol{x}_{\mathrm{c}}^{n-1}\} = \{X_n = 0|\boldsymbol{X}^{n-1} = \boldsymbol{x}_{\mathrm{c}}^{n-1}\}$ is either 0 or 1. This, in turn, implies that $\Lambda_{X_n|\boldsymbol{X}^{n-1}}(X_n|\boldsymbol{x}_{\mathrm{c}}^{n-1}) = 0$ almost surely, and therefore, conditioned on the event $\{\boldsymbol{X}^{n-1} = \boldsymbol{x}_{\mathrm{c}}^{n-1}\}$ one has that

$$\tau_m(\boldsymbol{X}^m) = \tau_n(\boldsymbol{x}_{\mathrm{c}}^{n-1}) \qquad \text{for all } m \geq n. \tag{7}$$

Finally, by using (4), one can show that

$$\mathbb{P}_w\left\{\pi_m = 0|\boldsymbol{X}^m, \boldsymbol{X}^{n-1} = \boldsymbol{x}_{\mathrm{c}}^{n-1}\right\} = F_w(\tau_m(\boldsymbol{X}^m)) = F_w(\boldsymbol{x}_{\mathrm{c}}^{n-1}) \tag{8}$$

Therefore, $\mathbb{P}_w\{\pi_m = 0|\boldsymbol{X}^{n-1} = \boldsymbol{x}_{\mathrm{c}}^{n-1}, X_n, \ldots, X_m\}$ is also either zero or one, showing that the m-th agent also is affected by a local information cascade.

□

Let us now introduce $U_s = \operatorname{ess\,sup} \Lambda_S(S_n)$, $U_{\xi_n} = \operatorname{ess\,sup} \xi_n$, $L_s = \operatorname{ess\,inf} \Lambda_S(S_n)$ and $L_{\xi_n} = \operatorname{ess\,inf} \xi_n$ as shorthand notations for the essential supermum and infimum of $\Lambda_S(S_n)$ and ξ_n [§]. In particular, U_s and L_s correspond to the signals within $\mathcal{S}$ that most strongly support the hypothesis $\{W = 1\}$ and $\{W = 0\}$, respectively. If one of these quantities diverge, this implies that there are signals $s \in \mathcal{S}$ that provide overwhelming evidence in favour of one of the competing hypotheses. On the other hand, if U_s and L_s are both finite then the agents are said to have *bounded beliefs* [3]. Similarly, when both U_ξ and L_ξ are finite we say agents have *bounded diversity*, which implies that the diversity among priors and cost functions is not too high. Using this notions, we present the main result of this work.

Theorem 1. *Local information cascades cannot take place when agents have either unbounded beliefs or unbounded diversity.*

Proof. Note first that, due to the independency between $\Lambda_S(S_n)$ and ξ_n, one has that

$$U_{\text{total}} \coloneqq \operatorname{ess\,sup}\{\Lambda_S(S_n) + \xi_n\} = U_s + U_\xi, \tag{9}$$

$$L_{\text{total}} \coloneqq \operatorname{ess\,inf}\{\Lambda_S(S_n) + \xi_n\} = L_s + L_\xi. \tag{10}$$

From this, U_{total} and L_{total} are unbounded if and only the agents have unbounded beliefs or unbounded diversity.

From Lemma 1, it is clear that π_n is fully determined by $\boldsymbol{x}^{n-1} \in \{0,1\}^{n-1}$ if and only if $\tau_n(\boldsymbol{x}^{n-1})$ is such that $F_w(\tau_n(\boldsymbol{x}^{n-1}))$ is zero or 1 for $w \in \{0,1\}$. Because of the definition of F_w, this happens whenever $\tau_n(\boldsymbol{X}^{n-1}) \notin [L_{\text{total}}, U_{\text{total}}]$, proving the Proposition. □

[§]The essential supremum is the smallest upper bound of a random variable that holds almost surely, being the natural measure-theoretic extension of the notion of supremum [20].

Information cascades are known to degrade the learning process, preventing the error rate of the learning process $\mathbb{P}\{\pi_n \neq W\}$ from converging to zero when the social network grows [4]. Therefore, Theorem 1 reveals a non-intuitive value of social diversity, as it can safeguard social learning from information cascades. In this way, social diversity can guarantee perfect social learning to happen asymptotically, even when agents have bounded beliefs and are hence prone to herd behaviour [4]. However, this benefit usually comes at the price of a slower convergence, which can be detrimental for the first agents of the decision sequence. This trade-off is explored in the next section.

4 Proof of concept

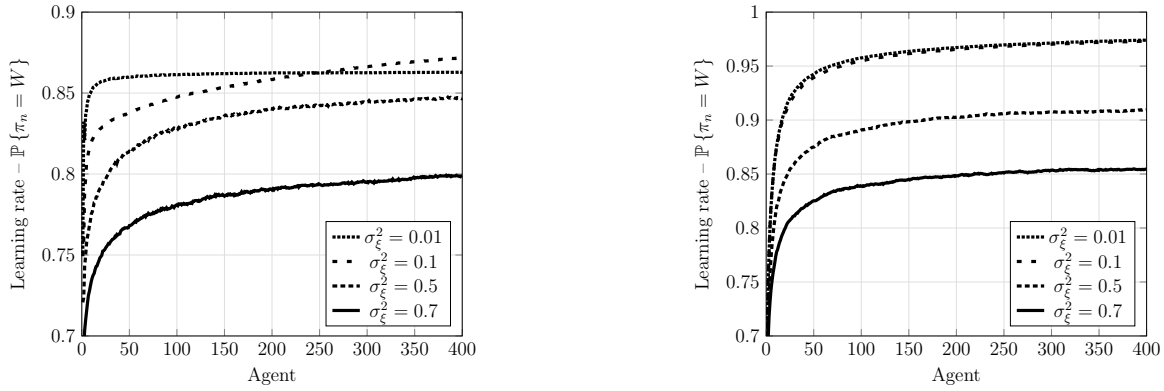
For illustrating the findings presented in Section 3, this section presents results of simulations of a social network following the model presented in Section 2. We considered two scenarios: one where S_n are binary variables that follow a binnary symmetric channel with $\mathbb{P}\{S_n \neq w|W = w\} = 1/4$, and other where S_n given $\{W = w\}$ are Gaussian variables $N(\mu_w, \sigma^2)$ with $\mu_w = (-1)^{1-w}$ and $\sigma^2 = 4$. These two signal models were choosen because it is known that agent following binary signals are strongly affected by information cascades, while agents following Gaussian signals are not affected by them (for further details about these scenarios c.f. [4, Section VI]). For simpicity, the social diversity has been modeled considering ξ_n to be i.i.d. following a Gaussian distribution $N(0, \sigma_\xi^2)$, and hence σ_ξ^2 quantifies the "diverstiy strength" of the social network. Each scenario was simulated 10^5 realizations, and the statistics of the learning error rate, defined as $\mathbb{P}\{\pi_n \neq W\}$ were computed afterwards.

In agreement with Theorem 1, results confirm that social learning processes can be benefited by social diversity. Figure 2a shows how the results of a collective inference carried out by agents driven by binary private signals achieve better performance asymptotically. However, for some values of social diversity the learning rate can be rather slow, making social learning not useful for small social networks. In all the studied cases it was seen that social diversity degrades the performance of the first agents in the decision sequence; however an adequate level of diversity can introduce a fast learning rate. In contrast, as illustrated in Figure 2a for agents following Gaussian signals, social diversity was found to be always detrimental in cases where agents have unbounded beliefs. This confirms the fact that the benefits of social diversity is to avoid information cascades, which are the main cause of poor performance of social learning in large networks [4].

The different effect that social diversity has over agents located at different positions in the inference process is further illustrated by Figure 3. We found that, for each agent, there exists an optimal level of social diversity that reduces the effect of information cascades without introducing too much noise. Agents located in the first places of the decision sequence are always affected negatively by social diversity, and hence for them is optimal to have $\sigma_\xi^2 = 0$.

5 Conclusion

This paper aims to undestand how social learning is affected when it is pursued by a diverse population. Our scenario considered rational agents with heterogeneous preferences, as encoded by their utility functions, and diverse prior information about the target variable. A communication theoretic analysis showed that this kind of social diversity is equivalent to additive noise in a communication channel. However, it was found that an unbounded social diversity prevent information cascades and, hence, introduces important improvements into the asymptotic social learning rate that can



(a) Binary private signals

(b) Gaussian private signals

Figure 2: Social learning rate for agents following binary or Gaussian private signals, under various levels of social diversity (σ_ξ^2). Social networks that follow binary signals are vulnerable to information cascades, and hence a non-zero social diversity improve their asymptotic learning rate. In contrast, social networks that follow Gaussian signals are inmune to information cascades, and hence social diversity have a purely detrimental effect.

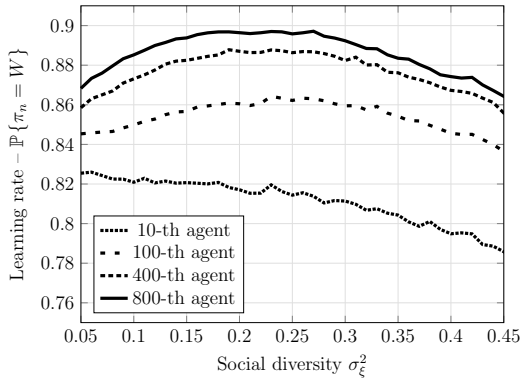


Figure 3: An optimal level of social diversity exist that can improve the social learning performance of agents located late in the inference process. However desirable, this better performance of late agents comes at the expense of a detrimental effect to the first agents.

be achieved by a population. Social learning is, therefore, one of those rare cases where noise can improve the overall performance.

To understand how can noise be beneficial, let us point out that rational social agents maximize their individual performance while ignoring the consequences of their actions on the aggregated behaviour. This selfish quality of the agent's behavior makes their actions locally optimal while being globaly suboptimal. In this context, the heterogeneity introduced by social diversity makes the decisions of each agent less informative to others, which reduces the information provided by the social network. This generates a reduced social pressure that, in turn, prevents information cascades and herd behaviour and introduces great improvements in the asymptotic social learning performance.

The benefits of social diversity are only experienced by agents that are prone to information cascades. Therefore, social diversity is not useful e.g. for agents with unbounded beliefs. However, in most applications agent's beliefs are bounded, either because their signals information content is limited or because the signals themselves

are bounded. The latter is the case in most engineering applications, e.g. the scenario studied in [10].

Finally, it is important to remark that social diversity provides benefits to the latter agents in the decision sequence, while degrading the performance of the first agents. Therefore, social diversity might in general be detrimental for the performance of social learning in small networks.

Acknowledgements

Fernando Rosas is supported by the European Union's H2020 research and innovation programme, under the Marie Skłodowska-Curie grant agreement No. 702981.

References

[1] Y. Bar-Yam, "Complexity rising: From human beings to human civilization, a complexity profile," *Encyclopedia of Life Support Systems (EOLSS), UNESCO, EOLSS Publishers, Oxford, UK*, 2002.

[2] D. Easley and J. Kleinberg, "Networks, crowds, and markets," *Cambridge University Press*, vol. 1, no. 2.1, pp. 2–1, 2010.

[3] D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian learning in social networks," *The Review of Economic Studies*, vol. 78, no. 4, pp. 1201–1236, 2011.

[4] F. Rosas, J.-H. Hsiao, and K.-C. Chen, "A technological perspective on information cascades via social learning," *IEEE Access*, vol. 5, pp. 22 605–22 633, 2017.

[5] J. Hsiao and K. C. Chen, "Steering information cascades in a social system by selective rewiring and incentive seeding," in *to be included in 2016 IEEE International Conference on Communications (ICC)*, 2016.

[6] L. Howell, "Digital wildfires in a hyperconnected world," *WEF Report*, 2013.

[7] A. V. Banerjee, "A simple model of herd behavior," *The Quarterly Journal of Economics*, pp. 797–817, 1992.

[8] S. Bikhchandani, D. Hirshleifer, and I. Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of political Economy*, pp. 992–1026, 1992.

[9] ——, "Learning from the behavior of others: Conformity, fads, and informational cascades," *The Journal of Economic Perspectives*, vol. 12, no. 3, pp. 151–170, 1998.

[10] F. Rosas and K.-C. Chen, "Social learning against data falsification in sensor networks," in *International Workshop on Complex Networks and their Applications*. Springer, 2017, pp. 704–716.

[11] F. Rosas, K.-C. Chen, and D. Gunduz, "Social learning for resilient data fusion against data falsification attacks," *arXiv preprint arXiv:1804.00356*, 2018.

[12] J. Mathiesen, N. Mitarai, K. Sneppen, and A. Trusina, "Ecosystems with mutually exclusive interactions self-organize to a state of high diversity," *Physical review letters*, vol. 107, no. 18, p. 188101, 2011.

[13] F. C. Santos, M. D. Santos, and J. M. Pacheco, "Social diversity promotes the emergence of cooperation in public goods games," *Nature*, vol. 454, no. 7201, p. 213, 2008.

[14] L. Smith and P. Sørensen, "Pathological outcomes of observational learning," *Econometrica*, vol. 68, no. 2, pp. 371–398, 2000.

[15] V. Bala and S. Goyal, "Conformism and diversity under social learning," *Economic theory*, vol. 17, no. 1, pp. 101–120, 2001.

[16] M. Loeve, *Probability Theory I.* Springer, 1978.

[17] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis.* CRC press Boca Raton, FL, 2014, vol. 2.

[18] T. M. Cover and J. A. Thomas, *Elements of information theory.* John Wiley & Sons, 2012.

[19] F. Rosas, V. Ntranos, C. J. Ellison, S. Pollin, and M. Verhelst, "Understanding interdependency through complex information sharing," *Entropy*, vol. 18, no. 2, p. 38, 2016.

[20] J. Dieudonne, *Treatise on Analysis.* Associated Press, New York, 1976, vol. II.

A Blockchain-based Signature Scheme for Dynamic Coalitions

Ricky A. Sewsingh[†], Jan C.A. van der Lubbe[†], Merel J. de Boer[‡]
[†]Delft University of Technology
Department of Intelligent Systems, Cyber Security Group, Faculty of EEMCS
P.O. Box 5031, 2600 GA Delft, The Netherlands
[‡]Water Authority Hollands Noorderkwartier
P.O. Box 250, 1700 AG Heerhugowaard, The Netherlands
R.A.Sewsingh@student.tudelft.nl, J.C.A.vanderLubbe@tudelft.nl, m.deboer@hhnk.nl

Abstract

In a secure dynamic coalition environment, secure communication and information sharing is very important. The fact that the coalition is dynamic implies that members are joining and leaving the coalition. Furthermore, it is common that the members do not fully trust each other, removing the possibility of having trusted third parties. *Van der Lubbe et al.*[1] introduced a distributed (n, n) signature scheme for dynamic coalitions, based on the distributed El Gamal signature scheme of *Park and Kurosawa*[2], and which can be expanded to a $(n + 1, n + 1)$ signature scheme. *Van Elsas et al.*[3] improved this signature scheme by enabling also the signature scheme to be reduced to a $(n - 1, n - 1)$ signature scheme and by preventing backlogging by the usage of One-Way Accumulators. In this paper we propose a further modification of this digital signature scheme by replacing the backlogging prevention functionality by the new Blockchain technology. The ledger capabilities of the Blockchain model provide means to effectively prevent backlogging.

1 Introduction

During military or peacekeeping operations it is often the case that coalitions are formed between nations or instances to achieve common objectives by joint decision making and resource sharing. Often a coalition partner participates for a certain period of time and other nations or instances might want to join the coalition in a later stage, resulting in a dynamic environment. In practice the members in these coalitions are dissimilar with regards to their disposition, requiring the coalition to be set up in a decentralised manner. In order to facilitate such an environment, a secure decentralised network should be in place that enables decision making, access control to shared information and the ability to join and leave in an appropriate way. Enabling the coalition to put its signature on important decisions and to be able to provide certified public keys to enable access control, the network should include a signature scheme. A challenge for such a signature scheme is that it should not be able for previous compositions of the coalition to create valid signatures. In other words, backlogging should be prevented in the system. *Van der Lubbe et al.*[1] introduced a distributed (n, n) signature scheme for dynamic coalitions based on the distributed El Gamal signature scheme of *Park and Kurosawa*[2], which can be expanded to a $(n + 1, n + 1)$ signature scheme. The scheme prevents backlogging by using two Oblivious Third Parties. *Van Elsas et al.*[3] then improved this signature scheme by enabling the signature scheme to be reduced to a $(n-1, n-1)$ signature scheme and by replacing the Oblivious Third Parties by One-Way Accumulators, dismissing the need for Third Parties in general.

The currently popular *Blockchain* model[4] has the capability to perform as a *ledger*. The ledger characteristics, specifically the chain of blocks, the Merkle Tree model and a consensus mechanism, can replace the functionality that prevents backlogging. The aim of this paper is to introduce a (n, n) signature scheme that is based on the Blockchain model for a secure dynamic coalition environment.

2 Background Information

2.1 Problem Scenario

The scenario has a *military setting*. This means that there is a high probability that the system will be targeted by malicious people. Additionally, trust is a big issue. It is very likely that the collaborating members of the coalition are independent organisations or countries who do not fully trust each other. This means the system should also be *decentralised*, where each member has the same influence and equal involvement. Lastly, because it is often the case that members of the coalition will leave and other entities might join the coalition, the coalition is *dynamic*.

The scenario involves *mission agreement*. Collaborating members are working together in a hostile environment where missions have to be executed. It is important that all members agree on a mission and its details. To ensure that this is the case, each mission is signed by all collaborating members, using a combined signature that is only valid if all members have cooperated in the generation of the signature. Signing ensures that every collaborating member agrees with the mission and therefore the mission can be executed safely. Also, if after signing a mission, a member claims it did not approve of this mission, the signature can be used to verify that the member in question did approve of the mission. Lastly, because in order to create a valid signature all members are needed, changes can not be made to a mission without the knowledge and agreement of all members.

3 The Blockchain-based Signature Scheme

The Blockchain-based signature scheme is comparable to the original signature scheme introduced by *Van Elsas et al*[3]. However certain parts of the protocol have been changed and the process has been altered to resemble a Blockchain model. This section is organised as follows. First, it is explained how consensus is achieved in the scheme. Secondly, the basic idea of the signature scheme is explained. Finally, a detailed description of the signature scheme, including the protocol descriptions, is provided. For convenience, throughout this section, the paper written by *Van Elsas et al.*[3] containing the original scheme will be referred to as the *original paper*.

3.1 Using the Blockchain Model

The Blockchain model, introduced by Nakamoto[4], has several characteristics which are useful in a secure dynamic coalition environment. The *chain of blocks* adds a layer of security, since changing a block also changes its subsequent blocks, thus increasing the difficulty of changing a block. The *Merkle Tree* can be used to store information efficiently, similar to the One-Way Accumulator[5], used in the original paper. However, *achieving consensus* is not so trivial. The original method of achieving consensus in the Bitcoin environment, the *Proof-of-Work* model, works, because users have an incentive, the transaction fee[6], to perform the work and blocks are created frequently. However, this does not hold in a secure dynamic coalition environment. Hence another method to achieve consensus has to be defined for this scheme.

The goal of the Proof-of-Work concept is to make a Block *immutable*, since changing a Block is explained to be hard. In this context this can also be achieved by *signing* a Block. If all members agree on the Block and its contents, the Block can be signed, ensuring the consensus of all members. This changes the Proof-of-Work to a new concept, *Proof-of-Consensus*.

3.2 Basic Idea

The first thing that needs to be noted is that because the Proof-of-Work is changed to Proof-of-Consensus, when a Block is signed, it results into a signature. A *Signed Block* is defined as the combination of the signature and the Block that has been signed. Note that the Signed Block replaces the *Hashed Block* from the original Blockchain model. Similarly to the original paper, the period between changes in the group composition is called a *phase*. A distinction is made between *regular signature issues*, signing regular messages, and *block signature issues*, signing Blocks.

During each phase, every member creates a Merkle Tree, containing all messages from regular signature issues. Due to the absence of a shared database, each member should define its own Merkle Tree. However, if the protocols are followed correctly by every member, the Merkle Trees will be identical. Whenever a regular signature is issued, each member adds the corresponding message to its Merkle Tree of that phase. In a similar way as the One-Way Accumulator in the original paper, this is used to prevent backlogging.

Once a change in the group composition is requested, the y value, the public key, of the new coalition is determined and the Block is defined. The Block contains the y value of the new coalition, the y value of the current coalition, the signature of the previously Signed Block and the Merkle Root from the current Merkle Tree.

Next, the Block is signed. This ensures that all members of the current coalition approve the change, hence ensuring a new member that all current members accept its admittance or ensuring a leaving member that all current members accept its departure. It also ensures that all members have knowledge and agree on the signatures that have been issued by the coalition. Signing the Block makes the Block immutable, since for a change to the Block, approval is needed by all members. And approval by all members makes the change valid. A Block contains the signature of the previously Signed Block. This functions as a pointer to the previous Block, hence creating a chain. If an earlier Block is changed, the corresponding signature is also changed, resulting in a change in all the subsequent Blocks.

The first Block, in the initialisation phase, only contains the y value of the coalition in the making. Since there was no previous coalition, there exists no previously Signed Block or Merkle Tree. This Block is then signed with the newly formed y value, to ensure that all members agree on forming a coalition.

Preventing Backlogging If a member finds a signature that is made with a y value that corresponds to a previous coalition and suspects backlogging, it only needs to find the Block that is signed using this y value and check if the message is contained in the Merkle Tree of that Block. If this is not the case, the signature is an instance of backlogging. It should be noted that every time a signature is being verified with a y value, which is not equal to the y value used by the latest coalition, this check needs to be done to prevent backlogging completely.

3.3 Detailed Description

This section describes the protocols of the signature scheme. The protocols are similar to the ones described in the original paper, however some changes are applied. First an overview of the definitions used in the protocols is provided. Then the five protocols

of the signature scheme are provided and elaborated. The five protocols discussed are the *Initialisation Protocol*, the *Joining Protocol*, the *Leaving Protocol*, the *Regular Signature Issue Protocol* and the *Block Signature Issue Protocol*.

3.3.1 Definitions

$\mathcal{N}$ is the current set of members of this coalition. The scheme enables members to leave and join this set. Next we have p and q which are large primes such that q divides $p-1$. g generates the group G_q which is a subgroup of $\mathcal{Z}_p$, of order q. We assume that p, q and g are publicly known. For the signature issue protocols, m is always the message that is going to be signed. The hash value of m is denoted by $h(m)$, where h is a publicly known hash function with a range from 1 to $q-1$. Lastly, the period between changes in the group composition is called a phase and is denoted by o.
A *Block* is denoted by $B_o(y_{o+1}, y_o, sig_{o-1}, Troot_o)$, where:

- $\diamond$ o = the phase the Block was created
- $\diamond$ y_{o+1} = the y value of the new composition going to be used in the $o+1th$-phase
- $\diamond$ y_o = the y value of the current composition that is used in the oth-phase
- $\diamond$ sig_{o-1} = the signature $sig_{o-1}(t, w, y)$ contained in the Signed Block SB_{o-1}, signed in the $o-1th$-phase
- $\diamond$ $Troot_o$ = the Merkle Root corresponding to the current Merkle Tree T_o that is used in the oth-phase.

Every phase o, each member defines a Merkle Tree T_o. All messages of the regular signature issues in phase o are stored in Merkle Tree T_o. When a group change occurs in phase o, Block B_o gets defined. First the new composition agrees on a new y value, that is going to be used in phase $o+1$, hence has the label y_{o+1}. Then the Block is defined, containing: the y value of the new composition, the y value of the current composition, the signature contained in the previously Signed Block and the Merkle Root corresponding to the Merkle Tree of the current phase. The Block has two functionalities. First of all, the Block contains the y value of the new composition, so when the Block is signed, all participating members are ensured that all members of the current coalition agreed on that value and therefore agreed on the new composition of the coalition. Secondly, the Block contains the Merkle Root of the Merkle Tree of the current phase. When the Block is signed, all coalition members are also ensured that the Merkle Tree is accepted, thus meaning all coalition members agree that the messages stored in the Merkle Tree are valid.
A *Signed Block* is denoted by $SB_o(B_o, sig_o(t, w, y))$, where:

1. o = the phase the Block was signed in
2. B_o = the Block that was defined in the oth-phase
3. $sig_o(t, w, y)$ = the signature in the oth-phase corresponding to Block B_o with signature values t, w and y

When a group change occurs in phase o, Block B_o gets defined. In order to achieve consensus and make the Block immutable, a block signature is issued with Block B_o as message. The resulting signature $sig_o(t, w, y)$ is proof that the Block has been signed. Hence the combination of $(B_o, sig_o(t, w, y))$ is the *Signed Block*. To verify the validity of Block B_o, one only needs to verify the validity of the corresponding signature $sig_o(t, w, y)$. Once the Signed Block SB_o is created, phase $o+1$ starts.

The first block in the chain is a special block. This is because there exist no previous phase, hence there are no previous blocks or values. This Block is denoted by $B_0(y_1)$. It only contains the y value the first composition of the coalition has agreed upon. Once this Block is signed, the first phase, phase 1, starts.

3.3.2 Initialisation Protocol

For initialisation, that occurs in phase 0, the following protocol will be applied:

1. Each member $i \in \mathcal{N}$ chooses a random secret x_i from $\mathcal{Z}_q$.
2. Each member $i \in \mathcal{N}$ broadcasts $z_i = g^{x_i} \bmod p$ to all other members.
3. Each member in $\mathcal{N}$ computes $y_1 = \prod_{i \in \mathcal{N}} z_i = g^x \bmod p$.
4. The Block $B_0 := B_0(y_1)$ is defined.
5. A block signature is issued by the members in $\mathcal{N}$ with $m := (B_0)$, resulting in the Signed Block $SB_0 := SB_0(B_0, sig_0(t, w, y_1))$.
6. Each member in $\mathcal{N}$ defines a new empty Merkle Tree T_1.
7. The first phase starts with Merkle Tree T_1 and Signed Block SB_0, containing the approved y_1 value.

The shared secret is defined as follows: $x \triangleq \sum_{i \in \mathcal{N}} x_i$. The shared secret is not known by any member. If a member joins or leaves the coalition, the shared secret changes according to the individual shares of the secret key of the new composition of the coalition.

The initialisation protocol is to start up the system. This happens in phase 0 and is done by creating and signing the first Block B_0. First the members in $\mathcal{N}$ decide on a y value. This is stored in the first Block B_0. Then the Block is signed using the block signature protocol resulting in the Signed Block SB_0. Lastly, the first Merkle Tree T_1 is defined by each member. The first phase then starts with Merkle Tree T_1 and the Signed Block SB_0, containing the y_1 value. Notice that usually the y value in the signature of a Signed Block is the y value of the current phase, but in this special case, where there is no y value yet, the Block is signed with the upcoming y_1 value.

3.3.3 Joining Protocol

When a new member k wants to join the group in the current phase o, the following protocol will be applied. The extended group $\mathcal{N} \cup \{k\}$ is denoted by $\mathcal{N}'$.

1. k requests the Signed Blocks $SB_0, ..., SB_{o-1}$ from a member $i \in \mathcal{N}$.
2. The member i sends the Signed Blocks $SB_0, ..., SB_{o-1}$.
3. k verifies that the Signed Blocks $SB_0, ..., SB_{o-1}$ contain valid signatures and that the chain is build correctly.
4. Each member $i \in \mathcal{N}$ sends $z_i = g^{x_i} \bmod p$ to k.
5. k chooses a random x_k from $\mathcal{Z}_q$.
6. k broadcasts $z_k = g^{x_k} \bmod p$ to each member $i \in \mathcal{N}$.
7. Each member in $\mathcal{N}'$ computes $y_{o+1} = \prod_{i \in \mathcal{N}'} z_i$.
8. The Block $B_o := B_o(y_{o+1}, y_o, sig_{o-1}, Troot_o)$ is defined and a block signature is issued by the members in $\mathcal{N}$ with $m := (B_o)$ resulting in a new Signed Block $SB_o := SB_o(B_o, sig_o(t, w, y_o))$.
9. A member $i \in \mathcal{N}$ sends SB_o to k.
10. k checks if SB_o is valid by verifying the validity of the signature of SB_o and if sig_{o-1} contained in Block B_o equals the signature of Signed Block SB_{o-1}.
11. Each member in $\mathcal{N}'$ defines a new empty Merkle Tree T_{o+1}.

12. The new phase $o+1$ starts with the group $\mathcal{N}' : \mathcal{N} := \mathcal{N}'$, Merkle Tree T_{o+1} and Signed Block SB_o, containing the approved y_{o+1} value.

When a new member k requests an admittance to the coalition in phase o, the joining protocol is issued. The member k requests the Signed Blocks $SB_0, ..., SB_{o-1}$. Upon retrieval, k verifies the validity of the signatures of the Signed Blocks and that the chain is build in a correct manner. Then the new y_{o+1} value is determined by the members of the new coalition, which includes the new member k. The Block B_o is then defined, containing the new y_{o+1} value, the currently used y_o value, the signature sig_{o-1} of the previously Signed Block SB_{o-1} and the Merkle Root $Troot_o$ of the current Merkle Tree T_o. This Block is signed using the block signature protocol, resulting in the Signed Block $SB_o(B_o, sig_o(t, w, y_o))$. The newly Signed Block is send to k, which then verifies the validity of its signature and checks if the signature of the previously Signed Block in Block B_o is correct. Lastly, every member in the new coalition $\mathcal{N}'$ defines a new empty Merkle Tree T_{o+1}. Now the new phase $o+1$ starts with Merkle Tree T_{o+1} and the Signed Block SB_o, containing the y_{o+1} value.

Verifying the Validity of the Chain As earlier stated, upon retrieval, member k verifies the validity of the signatures of the Signed Blocks and that the chain is build in a correct manner. This is done to ensure member k that the coalition has been performing the group change protocols correctly and no malicious attempt has been made to change the blocks in an incorrect manner.

This is done as follows. Starting with Signed Block SB_0, it is checked if SB_0 contains a valid signature. For all subsequent Signed Blocks SB_i, it is first checked if the signature of the previously Signed Block contained in Block B_i, sig_{i-1}, actually matches the signature of Signed Block SB_{i-1}. This ensures that the chain is build correctly. Then it is checked if the Signed Block SB_i contains a valid signature.
The validity of the signature $sig_o(t, w, y)$ is verified by

$$w \equiv (g^{t/h(m)} y^{-w/h(m)} \bmod p) \bmod q$$

3.3.4 Leaving Protocol

When a member j wants to leave the group in the current phase o, the following protocol will be applied. The reduced group $\mathcal{N} \backslash \{j\}$ is denoted by $\mathcal{N}'$.

1. Each member in $\mathcal{N}'$ computes $y_{o+1} = \prod_{i \in \mathcal{N}'} z_i$.
2. The Block $B_o := B_o(y_{o+1}, y_o, sig_{o-1}, Troot_o)$ is defined and a block signature is issued by the members in $\mathcal{N}$ with $m := (B_o)$ resulting in a new Signed Block $SB_o := SB_o(B_o, sig_o(t, w, y_o))$.
3. Each member in $\mathcal{N}'$ defines a new empty Merkle Tree T_{o+1}.
4. The new phase $o+1$ starts with the group $\mathcal{N}' : \mathcal{N} := \mathcal{N}'$, Merkle Tree T_{o+1} and Signed Block SB_o, containing the approved y_{o+1} value.

When a member j wishes to leave the coalition in phase o, the leaving protocol is issued. First the new y_{o+1} value is determined by the members of the new coalition, without the member j. Since every member already has the individual z_i values of the members, this is a trivial task and no communication is needed. Then the Block B_o is defined, containing the new y_{o+1} value, the currently used y_o value, the signature sig_{o-1} of the previously Signed Block SB_{o-1} and the Merkle Root $Troot_o$ of the current Merkle Tree T_o. This Block is signed using the block signature protocol, resulting in the Signed Block $SB_o(B_o, sig_o(t, w, y_o))$. The Block is signed with the members of $\mathcal{N}$, ensuring that the member j also agrees on the departure. Lastly, every member of the new coalition $\mathcal{N}'$ defines a new empty Merkle Tree T_{o+1}. The new phase $o+1$ starts with Merkle Tree T_{o+1} and the Signed Block SB_o, containing the y_{o+1} value.

3.3.5 Regular Signature Issue Protocol

When a regular signature is issued in the current phase o, the following protocol is applied.

1. Each $i \in \mathcal{N}$ chooses a random β_i from $\mathcal{Z}_q$.

$$\beta \triangleq \sum_{i \in \mathcal{N}} \beta_i$$

   Here β is the shared random secret, not known by any member.
2. Each $i \in \mathcal{N}$ broadcasts $c_i = g^{\beta_i} \bmod p$ to all other members.
3. Each $i \in \mathcal{N}$ reveals $a_i = g^{\gamma_i}$ where $\gamma_i \triangleq wx_i + h(m)\beta_i \bmod q$. Here w is equal to $v \bmod q$ with $v = \prod_{i \in \mathcal{N}} c_i = g^{\beta} \bmod p$.
4. Each member in $\mathcal{N}$ verifies that $\forall j, a_j = (z_j)^w (c_j)^{h(m)}$.
5. Each member in $\mathcal{N}$ computes $a = \prod_{i \in \mathcal{N}} a_i = \prod_{i \in \mathcal{N}} g^{\gamma_i} = g^{\sum_{i \in \mathcal{N}} \gamma_i} = g^t$ where $t = wx + h(m)\beta \bmod q$.
6. Each member in $\mathcal{N}$ adds m to its Merkle Tree T_o.

The validity of the signature $sig_o(t, w, y)$ is verified by

$$w \equiv (g^{t/h(m)} y^{-w/h(m)} \bmod p) \bmod q$$

When the coalition wishes to sign a message m with a regular signature in phase o, the regular signature protocol is issued. A shared random secret is created and accordingly the signature values can be created. If a valid signature for the message m is created, the message is added by each member of the coalition to its Merkle Tree T_o. It is possible to verify the validity of the signature using the formula provided.

3.3.6 Block Signature Issue Protocol

When a block signature is issued with message $m := (B_o)$ in the current phase o, the following protocol is applied.

1. Each $i \in \mathcal{N}$ chooses a random β_i from $\mathcal{Z}_q$.

$$\beta \triangleq \sum_{i \in \mathcal{N}} \beta_i$$

   Here β is the shared random secret, not known by any member.
2. Each $i \in \mathcal{N}$ broadcasts $c_i = g^{\beta_i} \bmod p$ to all other members.
3. Each $i \in \mathcal{N}$ reveals $a_i = g^{\gamma_i}$ where $\gamma_i \triangleq wx_i + h(m)\beta_i \bmod q$. Here w is equal to $v \bmod q$ with $v = \prod_{i \in \mathcal{N}} c_i = g^{\beta} \bmod p$.
4. Each member in $\mathcal{N}$ verifies that $\forall j, a_j = (z_j)^w (c_j)^{h(m)}$.
5. Each member in $\mathcal{N}$ computes $a = \prod_{i \in \mathcal{N}} a_i = \prod_{i \in \mathcal{N}} g^{\gamma_i} = g^{\sum_{i \in \mathcal{N}} \gamma_i} = g^t$ where $t = wx + h(m)\beta \bmod q$.
6. Each member in $\mathcal{N}$ defines the signature $sig_o(t, w, y)$ for message m.

7. Each member in $\mathcal{N}$ defines the Signed Block $SB_o(m, sig_o(t, w, y))$ with the message $m := (B_o)$.

The validity of the signature $sig_o(t, w, y)$ is verified by

$$w \equiv (g^{t/h(m)} y^{-w/h(m)} \bmod p) \bmod q$$

This protocol is very similar to the regular signature protocol. When a coalition change occurs in phase o, either the *joining protocol* or the *leaving protocol* is issued. These protocols include signing a Block B_o using the block signature issue protocol. This protocol is then issued with message $m := (B_o)$. A shared random secret is created and accordingly the signature values can be created. Now each member of the current coalition defines the signature $sig_o(t, w, y)$ that corresponds to the Block B_o and the Signed Block can be defined. Each member of the current coalition defines the new Signed Block $SB_o(m, sig_o(t, w, y))$ where $m := (B_o)$.

4 Discussion

4.1 Advantages

Because the memberlog used in the signature scheme of *Van Elsas et al*[3] has been replaced by the chain of blocks, maliciously changing information is more challenging. This is because when changing information, not only the corresponding block, but also all its subsequent blocks have to be re-signed.

The proposed signature scheme prevents backlogging by the use of Merkle Trees where *Van Elsas et al.*[3] uses a combination of One-Way Accumulators and a memberlog. The One-Way Accumulator is explained to be a combination of a hash function and a boolean array with a predefined size. A Merkle Tree does not require a predefined size and because only the Merkle Root of the Merkle Tree is stored in a Block, there is also an increase in efficiency regarding memory usage. Additionally, instead of using and storing the One-Way Accumulators and the memberlog separately, the proposed scheme stores all the information on the Blockchain. Each Block contains information about the composition, the public key y, and the regular messages, the Merkle Root. This makes the proposed scheme more efficient than the scheme in the original paper.

Bitcoin has a blocksize controversy[7]. Increasing the blocksize increases the throughput of the system, but decreases the fairness of mining a block. Large miners will have an advantage over small miners, resulting in centralisation of the mining power. This breaks the idea of having a decentralised system and results into a loss of mining power. However since the *Proof-of-Work* concept has been replaced by the *Proof-of-Consensus* concept, mining blocks is not applied, hence this is not an issue in this system. There is no limit to the blocksize, except the input limit of the hash function used in the *Block signature issue protocol*, to hash the Block.

4.2 Issues

It is possible for a set of members to change information in the Blockchain. If a block is chosen, where the set of members can create valid signatures for this block and its subsequent blocks, the members can change information in this block and re-sign the block and its subsequent blocks, changing information stored on Blockchain. This is possible if the set of members can form all compositions which were used to sign these blocks.

5 Future Work

Further research can be done to find out if the Blockchain scheme can be extended to perform several tasks, instead of one. *Verma et al.*[8] explained the idea of *Tear Sheets*, which can be implemented using the Blockchain model.

For this paper, the scenario is given where a coalition wants to sign messages regarding mission agreement. However this scenario can be extended where a coalition is able to provide *credentials*. This way the coalition can act like a distributed *Certificate Authority*, resulting in the possibility of creating a access control system.

Further research can be done to see if the Blockchain model can replace the complete signature scheme that is based on the signature scheme of *Park and Kurosawa*[2]. The purpose of a signature scheme can be seen as being able to create an immutable, unforgeable and verifiable connection between an identity and a document/message. By for example storing tuples of messages and verifiable identities on a Blockchain, the Blockchain can potentially provide the functionality of a signature scheme. By using a proper consensus mechanism, the information stored on the Blockchain can be seen as immutable. However, for it to be secure, a proper consensus mechanism has to be defined, verifiable identities have to be defined and used and the issue of unforgeability has to be addressed.

References

[1] J. C. A. van der Lubbe, M. J. de Boer, and Z. Erkin, "A signature scheme for a dynamic coalition defence environment without trusted third parties," in *International Conference on Cryptography and Information Security in the Balkans*. Springer, Oct 2014, pp. 237–249.

[2] C. Park and K. Kurosawa, "New elgamal type threshold digital signature scheme," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 79, no. 1, pp. 86–93, Jan 1996.

[3] M. Van Elsas, J. C. A. Van der Lubbe, and J. H. Weber, "A dynamic digital signature scheme without third parties," in *Proceedings of the 36th WIC Symposium on Information Theory in the Benelux and the 5th Joint WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux*, May 2015, pp. 89–95.

[4] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," https://bitcoin.org/bitcoin.pdf, 2008.

[5] J. Benaloh and M. De Mare, "One-way accumulators: A decentralized alternative to digital signatures," in *Workshop on the Theory and Application of of Cryptographic Techniques*. Springer, May 1993, pp. 274–285.

[6] J. A. Kroll, I. C. Davey, and E. W. Felten, "The economics of bitcoin mining, or bitcoin in the presence of adversaries," in *Proceedings of WEIS*, vol. 2013, June 2013, pp. 11–32.

[7] I. Eyal, A. E. Gencer, E. G. Sirer, and R. Van Renesse, "Bitcoin-ng: A scalable blockchain protocol." in *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, Mar 2016, pp. 45–59.

[8] D. Verma, N. Desai, A. Preece, and I. Taylor, "A blockchain based architecture for asset management in coalition operations," in *SPIE Defense+ Security*. International Society for Optics and Photonics, May 2017, pp. 101 900Y–101 900Y.

Fingerprint template protection with spectral minutia-pair representations

Taras Stanko, Bin Chen, Boris Škorić
TU Eindhoven {t.stanko, b.c.chen, b.skoric}@tue.nl

Helper Data Systems (HDSs) are a cryptographic primitive that allows for the reproducible extraction of secrets from noisy measurements, such as biometrics. A HDS typically requires the biometric template to have a fixed-length representation. Recently a pair-based spectral minutiae approach was introduced [1] to obtain such a representation. We construct a HDS based on minutia-pair spectral functions.

Minutiae are special points in a fingerprint. Let Z be the number of detected minutiae. Let (x_a, y_a) be the coordinates of the a'th minutia and θ_a its orientation. Let $R_{ab} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$, $\tan\phi_{ab} = \frac{y_a - y_b}{x_a - x_b}$. In [1] a spectral function was introduced,

$$M_{x\theta}(q, R) = \sum_{a\in\{1,\ldots,Z\}} \sum_{b>a} e^{iq\phi_{ab}} e^{-\frac{(R-R_{ab})^2}{2\sigma^2}} e^{i(\theta_b - \theta_a)}, \quad (1)$$

evaluated on a grid $R \in \{16, 22, \ldots, 130\}$, $q \in \{1, 2, \ldots, 16\}$. Taking pairs ensures translation invariance. We now introduce a second orientation-dependent spectral function, which captures information complementary to $M_{x\theta}$,

$$M_{x\beta}(q, R) = \sum_{a\in\{1,\ldots,Z\}} \sum_{b>a} e^{i(q-2)\phi_{ab} + i(\theta_b + \theta_a)} e^{-\frac{(R-R_{ab})^2}{2\sigma^2}}. \quad (2)$$

We construct a HDS based on $M_{x\theta}$ and $M_{x\beta}$, making use of Zero-Leakage quantisation [2] and the Code Offset Method. We test out HDS on public databases, MCYT [3], FVC2000 and FVC2002. We show ROC curves at various stages of the data processing: in the analog domain, after naive quantisation, after Zero Leakage quantisation, with/without application of an error-correcting code and with/without selection of reliable components. The best results are obtained when we average the spectral functions from multiple enrollment images, apply Zero Leakage quantisation, and select 1024 reliable components. (These components are not person-dependent and hence do not leak.) The bit error rates are high, 0.23 to 0.34. Hence we have used Polar codes, which perform well even at short codeword length. For the MCYT database (high-quality fingerprints) we obtain an Equal Error Rate of $\approx 1\%$ with an extracted message length of 25 bits. For the FVC2000 database we obtain an EER of $\approx 6\%$ with message length 15. For more details see the full version [4].

References

[1] T. Stanko, B. Škorić. *Minutia-pair spectral representations for fingerprint template protection.* IEEE Workshop on Information Forensics and Security (WIFS) 2017.
[2] T. Stanko, F.N. Andini, B. Škorić. *Optimized quantization in Zero Leakage Helper Data Systems.* IEEE Transactions on Information Forensics and Security, Vol.12, No.8, pp.1957–1966 (2017).
[3] J. Ortega-Garcia et al. *MCYT baseline corpus: A bimodal biometric database.* Vision, Image and Signal Processing, special issue on biometrics on the Internet, Vol.150, pp.395–401, IEEE (2003).
[4] T. Stanko, B.C. Chen, B. Škorić. *Fingerprint template protection using minutia-pair spectral representations.* http://arxiv.org/abs/1804.01744

PUF-Enabled Asymmetric Cryptography

R. Uppu, T.A.W. Wolterink, S.A. Goorden, B.C. Chen, B. Škorić, A.P. Mosk, P.W.H. Pinkse

We introduce a new scheme, based on optical PUFs and quantum optics, to achieve public-key encryption. The public key is a description of the PUF's challenge-response bahaviour. The private key is having-control-over-the-physical-object-itself. The security is based on a single technological assumption: the difficulty of performing a specific measurement on quantum states that distinguishes between several orthogonal subspaces. This work is a further development of Quantum-Secure Authentication (QSA) [1, 2, 3], with two main differences, (i) light has to travel only from Alice to Bob instead of both ways, (ii) a different security assumption. We call our scheme PUF-Enabled Asymmetric Communication (PEAC). Lab experiments show the feasibility of implementing PEAC in practice. Variants of PEAC able to handle a large number of erasures can be thought of as Quantum Key Distribution with PUF-based authentication of the recipient.

Bob has an optical PUF. Positioned behind it are detectors D_0 and D_1 which each cover half of the output space. Alice wants to send a message x. She encodes x in an ECC codeword $c \in \{0,1\}^n$. For each bit $i \in \{1,\ldots,n\}$ the following steps are executed. Alice prepares N photons in an identical quantum state $|\psi_i\rangle$, chosen randomly from $\mathcal{H}_{c_i}$, and sends them to Bob. Here $\mathcal{H}_0$ and $\mathcal{H}_1$ are subspaces of the K-mode Hilbert space (think of a multimode fiber with K modes), such that states in $\mathcal{H}_b$ will hit D_b after passing through the PUF. Bob detects a signal in D_0 or D_1.

The quantum channel is noisy. Bob receives $c' \in \{0,1\}^n$ and decodes this to $\hat{x}$, which should equal x. The idea is that the owner of the PUF is the only one who can reconstruct x from the received quantum states. The security assumption is that it is hard to do a measurement which can determine, more efficiently than Bob's PUF, the bit b from a state $|\psi\rangle^{\otimes N}$ with $|\psi\rangle \in \mathcal{H}_b$. Under this assumption, it can be shown that there is a positive secrecy rate as long as $N/(N+K) < P_{\text{Bob}} - 1/2$. ($P_{\text{Bob}}$ is Bob's probability of correctly distinguishing a bit.)

We did experiments on 13μm zinc oxide PUFs, in transmission, at 790 nm wavelength, over a distance of 2 meters, with $K = 900$, $N = 33$. The resulting channel has 59% erasure rate, and 43% bit error rate in the arriving bits. The noise is corrected using a Polar code with $n = 2^{15}$ and code rate 0.002, designed to eliminate Alice-to-Eve polarised channels. The decoder is a CRC-aided Successive Cancellation-List decoder. The frame error rate is below 10^{-4}. For details we refer to the full paper [4].

References

[1] B. Škorić. *Quantum readout of Physical Unclonable Functions.* Africacrypt 2010, LNCS 6055, pp.369–386.

[2] S.A. Goorden, M. Horstmann, A.P. Mosk, B. Škorić, P.W.H. Pinkse. *Quantum-secure authentication of a physical unclonable key.* Optica Vol.1 No.6, pp.421–424 (2014).

[3] B. Škorić, P.W.H. Pinkse, A.P. Mosk. *Authenticated communication from quantum readout of PUFs.* Quantum Information Processing, Vol.16 No.8, pp.200 (2017).

[4] R. Uppu, T.A.W. Wolterink, S.A. Goorden, B.C. Chen, B. Škorić, A.P. Mosk, P.W.H. Pinkse. *Asymmetric Cryptography with Physical Unclonable Keys.* `http://export.arxiv.org/abs/1802.07573`

The coset leader weight enumerator of the product code $[m, m-1, 2]_q \otimes [n, n-1, 2]_q$

Putranto Utomo[1,2] Ruud Pellikaan[1]
p.h.utomo@tue.nl g.r.pellikaan@tue.nl
[1]Eindhoven University of Technology
Dept. Mathematics and Computer Science
P.O. Box 513. 5600 MB Eindhoven
[2]Sebelas Maret University
Fac. of Mathematics and Natural Sciences

Abstract

One way to get a new code using existing codes is by means of a tensor product or Kronecker product, and the result is often simply called a product code. The idea is as follow. Let C and D be two linear $[n_1, k_1, d_1]$ and $[n_2, k_2, d_2]$ codes over the finite field of size q, respectively. Then, every codeword is written in a matrix form of size $n_1 \times n_2$ where each column belongs to C and each row belongs to D. Then, the product code $C \otimes D$ is a code of length $n_1 n_2$, dimension $k_1 k_2$ and minimum distance $d_1 d_2$.

The coset leader weight enumerator is one of the important aspects of an error correction code. For example, one can express the error probability of a code in term of its coset leader weight enumerator. However, finding the coset leader weight enumerator of an arbitrary code turns out to be very difficult.

Let C_n be the linear $[n, n-1, 2]_q$ code with parity check matrix the all ones vector. In this paper, we present a closed formula in the binary and ternary case of the coset leader weight enumerator of the product code $C_m \otimes C_n$ for arbitrary m and n.

1 Introduction

Despite of the difficulties of finding the coset leader weight enumerator of a code [3], it is one of the important aspects of an error correcting code, since for example one can represent the error probability of a code in term of its coset leader enumerator [4]. Moreover, it also plays an important role in steganography [2]. In [1], the authors address the problem of (extended) coset leader enumerator.

Let C be a code in $\mathbb{F}_q^n$ and $\mathbf{x}$ be any word in $\mathbb{F}_q^n$. The weight of the coset $\mathbf{x} + C$ is the minimal weight of an element in that coset, that is $\min\{\mathrm{wt}(\mathbf{x} + \mathbf{c} | \mathbf{c} \in C\}$. The coset leader of $\mathbf{x} + C$ is defined as a choice of one element in the coset $\mathbf{x} + C$ with minimal weight.

Now, let H be a parity check matrix for code C, that is a matrix where C is the null space of H. Suppose that $\mathbf{x}$ be a word in $\mathbb{F}_q^n$. Then, $H\mathbf{x}^T$ is defined as the syndrome for $\mathbf{x}$.

Remark also that the syndrome of $\mathbf{x}$ is in one-to-one correspondence to $\mathbf{x}+C$, since two words are in the same coset if and only if they have the same syndrome. Then $\mathrm{wt}(x + C)$ is the minimum number of columns of H such that a linear combination of these columns give $\mathbf{x}H^T$, that is the syndrome of $\mathbf{x}$ written as a column.

1.1 Coset leader weight enumerator

Let C be a linear code in $\mathbb{F}_q^n$. Then

$\alpha_i(C)$ is the number of coset leaders of C having weight i,

$\alpha_C(Z)$ is the coset leader weight enumerator of C and is given as follows:

$$\alpha_C(Z) = \sum_{i=0}^{n} \alpha_i(C) Z^i.$$

Recall that the covering radius $\rho(C)$ of the code C is defined as

$$\max\{d(\mathbf{x}, C) | \mathbf{x} \in \mathbb{F}_q^n\}.$$

In other word, we have that

$$\rho(C) = \max\{i | \alpha_i(C) \neq 0\}.$$

Remark that the covering radius may increase after extending the finite field. In case we want to stress that we consider the covering radius of the code C over $\mathbb{F}_q$, we denote it by $\rho(C, \mathbb{F}_q)$.

1.2 The $[n, n-1, 2]_q$ code

Let C_n be the $\mathbb{F}_q$-linear code of length n with parity check matrix $(1, 1, \cdots, 1)$. Then C_n has parameters $[n, n-1, 2]_q$. In the binary case C_n is the even weight code.

Remark that C_n has q cosets, one of weight 0 and $q-1$ of weight 1. Hence $\alpha_{C_n}(Z) = 1 + (q-1)Z$. Notice also that since this code has parity check matrix of size $1 \times n$, clearly that it has q different syndromes, which match with the number of cosets.

2 Product code

Let C be a code in $\mathbb{F}^m$ and D be a code in $\mathbb{F}^n$. Then, the product code $C \otimes D$ is a code with length mn in $\mathbb{F}_q^{m \times n}$, where if we write in a matrix form of size $m \times n$, all the columns are belong to C and all the rows are belong to D.

2.1 The product code $[m, m-1, 2]_q \otimes [n, n-1, 2]_q$

Let C_m be an $[m, m-1, 2]_q$ code and C_n be an $[n, n-1, 2]_q$ code. We consider the product code of C_m and C_n.

Definition 2.1 *Let* $\mathbf{x}$ *be an element in* $\mathbb{F}_2^{m \times n}$. *Define* $\mathbf{r}$ *in* $\mathbb{F}_q^m$ *and* $\mathbf{c}$ *in* $\mathbb{F}_q^n$ *as follows:*

$$r_i(\mathbf{x}) = r_i = \textit{syndrome of } i\textit{-th row of } \mathbf{x}, \textit{ that is: } \sum_{j=1}^{n} x_{ij},$$

$$c_j(\mathbf{x}) = c_j = \textit{syndrome of } j\textit{-th column of } \mathbf{x}, \textit{ that is: } \sum_{i=1}^{m} x_{ij}.$$

Lemma 2.2

$$\sum_{i=1}^{m} c_i(\mathbf{x}) = \sum_{j=1}^{n} r_j(\mathbf{x})$$

Proof. From the definition of $r_i(\mathbf{x})$ and $c_j(\mathbf{x})$, we have

$$\sum_{i=1}^{m} c_i = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m} x_{ij} = \sum_{j=1}^{n} r_j.$$

◇

Remark 2.3 *By the definition of the product code $C_m \otimes C_n$, we can construct a parity check matrix H for $C_m \otimes C_n$ of size $(m+n) \times mn$, where the first m rows are given by $r(\mathbf{x})$, the row syndromes of $\mathbf{x}$, and the last n rows are given by $c(\mathbf{x})$, the column syndromes of $\mathbf{x}$. But, since the last row of H is dependent on the previous one by Lemma 2.2, we can delete it and therefore obtain the parity check matrix H of size $(m+n-1) \times mn$. Below is an example of a parity check matrix of the product code $C_4 \otimes C_3$:*

$$\begin{pmatrix}
1 & 1 & 1 & 1 & & & & & & & & \\
 & & & & 1 & 1 & 1 & 1 & & & & \\
 & & & & & & & & 1 & 1 & 1 & 1 \\
1 & & & & 1 & & & & 1 & & & \\
 & 1 & & & & 1 & & & & 1 & & \\
 & & 1 & & & & 1 & & & & 1 & \\
\end{pmatrix}.$$

Proposition 2.4 *Let $\mathbf{x}$ be an element in $\mathbb{F}_q^{m\times n}$. Let $r_i = r_i(\mathbf{x})$ and $c_i = c_i(\mathbf{x})$. Let $s = \sum_{i=1}^{m} r_i$. Now, let $\mathbf{y}$ be the element in $\mathbb{F}_q^{m\times n}$ as given by the following matrix:*

y_{11}	c_2	c_3	$\cdots$	c_m
r_2	0	0	$\cdots$	0
r_3	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$
r_n	0	0	$\cdots$	0

where $y_{11} = r_1 + c_1 - s$. Then $\mathbf{x}$ and $\mathbf{y}$ have the same coset of $C_m \otimes C_n$ and

$$wt(\mathbf{y}) = \begin{cases} wt(c) + wt(r) - 2 & \text{if } y_{11} = 0 \\ wt(c) + wt(r) - 1 & \text{if } y_{11} \neq 0. \end{cases}$$

Proof. From the definition of $\mathbf{y}$, we have that

$$r_i(\mathbf{x}) = r_i(\mathbf{y}) \text{ for all } i,$$
$$c_j(\mathbf{x}) = c_j(\mathbf{y}) \text{ for all } j.$$

Hence $r_i(\mathbf{x}-\mathbf{y}) = 0$ for all i. So all rows of $\mathbf{x}-\mathbf{y}$ are in C_n. Similarly all columns of $\mathbf{x}-\mathbf{y}$ are in C_m. So $\mathbf{x}-\mathbf{y}$ is in $C_m \otimes C_n$. Hence $\mathbf{x}$ and $\mathbf{y}$ have the same coset of $C_m \otimes C_n$.

Notice that all entries in $\mathbf{y}$ are zeroes except in the first column and in the first row. Hence the weight of $\mathbf{y}$ is determined by the weight of the first column and the first row. Hence the weight of $\mathbf{y}$ is determined as stated in the lemma.

Remark that $s = \sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j$, and $y_{11} = r_1 + c_1 - s$ is uniquely determined by r_1, c_1, and s.

◇

Definition 2.5 *Let* $\mathbf{x}$ *in* $\mathbb{F}_q^{m\times n}$ *and* α, β *in* $\mathbb{F}_q$. *Then define* $I_\alpha(\mathbf{x})$ *and* $J_\beta(\mathbf{x})$ *as follows:*

$$I_\alpha(\mathbf{x}) = \{i \mid r_i = \alpha\},$$
$$J_\beta(\mathbf{x}) = \{j \mid r_j = \beta\}.$$

Lemma 2.6 *Let* $\mathbf{x}$ *in* $\mathbb{F}_q^{m\times n}$. *Then*

$$\{I_\alpha(\mathbf{x}) \mid \alpha \in \mathbb{F}_q\} \text{ is a partition of } \{1, \ldots, m\},$$
$$\{J_\beta(\mathbf{x}) \mid \beta \in \mathbb{F}_q\} \text{ is a partition of } \{1, \ldots, n\}.$$

Proof. Let $S = \{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ be the set of syndromes of the rows of $\mathbf{x}$. Since a row has a unique syndrome $r_i(\mathbf{x})$, then for any $\alpha_i \neq \alpha_j \in S$, $I_{\alpha_i}(\mathbf{x}) \cap I_{\alpha_j}(\mathbf{x}) = \emptyset$. Furthermore, every row has a syndrome, hence

$$\bigcup_{\alpha\in S} I_\alpha(\mathbf{x}) = \{1, \ldots, m\}.$$

Therefore, $\{I_\alpha(\mathbf{x}) \mid \alpha \in \mathbb{F}_q\}$ is a partition of $\{1, \ldots, m\}$. The proof is similar for the columns. ◇

Lemma 2.7 *Let* $\mathbf{x}$ *in* $\mathbb{F}_q^{m\times n}$. *Then the weight of the coset leader of* $\mathbf{x} + C_m \otimes C_n$ *is at least*

$$\max\left\{\sum_{\alpha\in\mathbb{F}_q^*} |I_\alpha(\mathbf{x})|, \sum_{\alpha\in\mathbb{F}_q^*} |J_\alpha(\mathbf{x})|\right\}.$$

Proof. Let α be an element in $\mathbb{F}_q^*$. Then, referring to Lemma 2.6, $I_\alpha(\mathbf{x})$ and $J_\alpha(\mathbf{x})$ are elements of the partition of $\{1, \ldots, m\}$ and $\{1, \ldots, n\}$, respectively.

Then, for every nonzero α there are at least $\max\{|I_\alpha(\mathbf{x})|, |J_\alpha(\mathbf{x})|\}$ positions that need to be changed so that their syndrome becomes 0. In other words, the weight of the coset leader of $\mathbf{x} + C_m \otimes C_n$ is at least $\max\{\sum_{\alpha\in F_q^*} |I_\alpha(\mathbf{x})|, \sum_{\alpha\in F_q^*} |J_\alpha(\mathbf{x})|\}$.

◇

2.2 The binary case

Now, let C_m and C_n be the binary even weight codes with length m and n, respectively. Let $\mathbf{x}$ in $\mathbb{F}_2^{m\times n}$. Then $I_1(\mathbf{x})$ and $J_1(\mathbf{x})$ are the only interesting sets in the binary case. So

$$I(\mathbf{x}) := I_1(\mathbf{x}) = \{i \mid \text{the number of ones in row } i \text{ of } \mathbf{x} \text{ is odd}\},$$
$$J(\mathbf{x}) := J_1(\mathbf{x}) = \{j \mid \text{the number of one in column } j \text{ of } \mathbf{x} \text{ is odd}\}.$$

Lemma 2.8 *We have that* $wt(\mathbf{x}) \equiv |I(\mathbf{x})| \equiv |J(\mathbf{x})| \mod 2$.

Proof. This follows from Lemma 2.2 for $q = 2$. An alternative proof is as follow: Let $\mathbf{x} \in \mathbb{F}_2^{m\times n}$. If we count the number of ones of $\mathbf{x}$ first row wise, then we see that the parities of $I(\mathbf{x})$ and $\mathrm{wt}(\mathbf{x})$ are equal. Hence $\mathrm{wt}(\mathbf{x}) \equiv |I(\mathbf{x})| \mod 2$. Similarly we get that $\mathrm{wt}(\mathbf{x}) \equiv |J(\mathbf{x})| \mod 2$. ◇

Lemma 2.9 *Every coset leader of* $\mathbf{x} + C_m \otimes C_n$ *is in one-to-one correspondence with a pair* (I, J) *such that* $I \subseteq \{1, \cdots, m\}, J \subseteq \{1, \cdots, n\}$ *and* $|I| \equiv |J| \mod 2$.

Proof. Let $\mathbf{x}$ and $\mathbf{y}$ in the $\mathbb{F}_q^{m\times n}$ with the same pair of syndromes such that $I(\mathbf{x}) = I(\mathbf{y})$ and $J(\mathbf{x}) = J(\mathbf{y})$. Then the number of ones in every row of $\mathbf{x} - \mathbf{y}$ is even, and also for every column. So $\mathbf{x} - \mathbf{y} \in C_m \otimes C_n$. Hence $\mathbf{x}$ and $\mathbf{y}$ are in the same coset.

Now, suppose we have subsets I and J of $\{1, \ldots, m\}$ and $\{1, \ldots, n\}$, respectively, where $|I| \equiv |J| \mod 2$. We may assume without loss of generality that $|I| \leq |J|$ and

$$I = \{i_1, i_2, \ldots, i_r\},$$
$$J = \{i_1, i_2, \ldots, i_s\}.$$

Let $\mathbf{e} \in \mathbb{F}_2^{m\times n}$ where $e_{i_k,j_k} = 1$ for $k = 1, \ldots, r$, and $e_{1,j_l} = 1$ for $l = r + 1, \ldots, s$, and zero elsewhere. Then $\mathbf{e}$ has the form as depicted in the following array.

$$\left[\begin{array}{ccc|ccc} 1 & & & 1 & 1 & \cdots & 1 \\ & \ddots & & & & & \\ & & 1 & & & & \end{array}\right]$$

Since $|I| \equiv |J| \mod 2$, then $s - r = 0 \mod 2$. So $I(\mathbf{e}) = I$ and $J(\mathbf{e}) = J$. ◇

Proposition 2.10 *The weight of the coset* $\mathbf{x}+C_m\otimes C_n$ *is equal to* $\max\{|I(\mathbf{x})|, |J(\mathbf{x})|\}$.

Proof. Suppose that $\mathbf{x}$ is any word in $\mathbb{F}_2^{m\times n}$, and

$$I(\mathbf{x}) = \{i_1, i_2, \cdots, i_r\},$$
$$J(\mathbf{x}) = \{j_1, j_2, \cdots, j_s\}.$$

Then $s \equiv r \mod 2$ by Lemma 2.8.

Without loss of generality, we may assume that $r \leq s$. Thus, by Lemma 2.7 for the case $q = 2$, the weight $\mathbf{x} + C_m \otimes C_n$ is at least equal to $s = \max\{I(\mathbf{x}), J(\mathbf{x})\}$.

Remark that the number of ones in each column and in each row does not change if we perform row and column permutations. Therefore, there exists an $\mathbf{e}$ as given in Lemma 2.9 such that $\mathrm{wt}(\mathbf{e}) = \mathrm{wt}(\mathbf{x})$. Hence $I(\mathbf{x}) = I(\mathbf{e})$ and $J(\mathbf{x}) = J(\mathbf{e})$. Then, $\mathbf{x}$ and $\mathbf{e}$ have the same coset by Lemma 2.9. Therefore the weight of the coset $\mathbf{x}+C_m\otimes C_n$ is equal to $s = \max\{I(\mathbf{x}), J(\mathbf{x})\}$.

◇

Corollary 2.11

$$\rho(C_m \otimes C_n, \mathbb{F}_2) = \max\{m, n\}$$

Proof. The result follows directly from the definition of the covering radius and Proposition 2.10. ◇

Proposition 2.12 *The number of coset leaders of the product code $C_m \otimes C_n$ of weight t is given as follows*

$$\alpha_t(C_m \otimes C_n) = \sum_{\substack{r \equiv t \bmod 2 \\ r \leq t}} \binom{m}{r}\binom{n}{t} + \sum_{\substack{s \equiv t \bmod 2 \\ s < t}} \binom{m}{t}\binom{n}{s}.$$

Proof. Without loss of generality, we will proof the first part of the equation. From the proof of Lemma 2.10, we see that the number of coset leader with weight t is the same as choosing the t columns from n, and choosing r rows under condition that $r \equiv t \bmod 2$. ◇

2.3 The ternary case

Suppose we have $\mathbf{x} \in \mathbb{F}_3^{m \times n}$ with the set of row syndromes $\{I_\alpha\}$ and column syndromes $\{J_\alpha\}$, where $\alpha \in \{0, 1, 2\}$. Since we are interested only in the non-zero syndrome, we only consider α equal to 1 and 2.

Lemma 2.13 *Let $x \in \mathbb{F}_3^{m \times n}$. Then*

$$\mathrm{wt}(x + C_n \otimes C_m) = \begin{cases} |J_1| + |J_2| & \text{if } |I_1| \leq |J_1| \text{ and } |I_2| \leq |J_2| \\ |I_1| + |I_2| & \text{if } |J_1| \leq |I_1| \text{ and } |J_2| \leq |I_2|. \end{cases}$$

Proof. The right hand side of the formula is a lower bound for the weight of the coset of $\mathbf{x}$ by Lemma 2.7. In the first case, we have $|J_1| + 2|J_2| \equiv |I_1| + 2|I_2|$ by Lemma 2.2, and therefore $|J_1| - |I_1| + 2|J_2| - 2|I_2| \equiv 0$. After a permutation of columns and rows, there exists an $\mathbf{e}$ in the coset of $\mathbf{x}$ of weight $|J_1| + |J_2|$ with minimum weight which has the following structure:

$1 \quad \ddots \quad 1$	$1 \quad \cdots \quad 1$		$2 \quad \cdots \quad 2$
		$2 \quad \ddots \quad 2$	

Therefore, the weight of $\mathbf{x} + C_m \otimes C_n$ is equal to $|J_1| + |J_2|$.

A similar argument holds in case $|I_1| \geq |J_1|$ and $|I_2| \geq |J_2|$. ◇

Lemma 2.14 *Let $x \in \mathbb{F}_3^{m \times n}$. Then*

$$\mathrm{wt}(x + C_n \otimes C_m) = \begin{cases} |J_1| & \text{if } |I_1| = |J_2| = 0 \text{ and } 2|I_2| \leq |J_1| \\ |J_2| & \text{if } |I_2| = |J_1| = 0 \text{ and } 2|I_1| \leq |J_2| \\ |I_2| & \text{if } |I_1| = |J_2| = 0 \text{ and } 2|J_1| \leq |I_2| \\ |I_1| & \text{if } |I_2| = |J_1| = 0 \text{ and } 2|J_2| \leq |I_1|. \end{cases}$$

Proof. Without loss of generality, we may assume in the first case that $|I_2| = m$ and $|J_1| = n$. From Lemma 2.2, we know that $\sum_{i=1}^{m} c_i(\mathbf{x}) \equiv \sum_{j=1}^{n} r_j(\mathbf{x})$, and hence $2|I_2| \equiv |I_1| \bmod 3$.

Suppose that $2|I_2| \leq |J_1|$. Remark that since $|J_1| > |J_2| = 0$, the minimum weight of the coset is at least n. Now we will show that there is an element in the coset of $\mathbf{x}$ of weight n. Since $|J_1| \geq 2|I_2|$ and $|J_1| \equiv 2|I_2| \bmod 3$, the following holds

$$|J_1| = 2|I_2| + 3k.$$

So $n = 2m + 3k$.

Now consider the ternary $m \times n$ matrix $\mathbf{e}$ defined by $\mathbf{e}_{i,2i-1} = 1$ and $\mathbf{e}_{i,2i} = 1$ for $1 \leq i \leq m$, and $\mathbf{e}_{m,j} = 1$ for $2m < j \leq n$, and $\mathbf{e}_{i,j} = 0$ otherwise, as depicted in the figure below:

$$\left[\begin{array}{cccccc|ccc} 1 & 1 & & & & & & & \\ & & 1 & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & 1 & 1 \cdots & 1 \end{array}\right]$$

Then every column has exactly one 1 and the first $m-1$ rows have exactly two 1's and the number of 1's in the last row is $2 + 3k$, $k \in \mathbb{Z}$. Hence $\mathbf{e}$ has weight n and has the same coset as $\mathbf{x}$.

A similar proof is given in second, third, and fourth case. ◇

Lemma 2.15 *Let $x \in \mathbb{F}_3^{m \times n}$. Then*

$$\mathrm{wt}(x + C_n \otimes C_m) = \begin{cases} \frac{2}{3}(|I_2| + |J_1|) & \text{if } I_1 = J_2 = \emptyset,\ |J_1| \leq 2|I_2| \text{ and } |I_2| \leq 2|J_1| \\ \frac{2}{3}(|I_1| + |J_2|) & \text{if } I_2 = J_1 = \emptyset,\ |J_2| \leq 2|I_1| \text{ and } |I_1| \leq 2|J_2|. \end{cases}$$

Proof. Let $\mathbf{e}$ be an element in $\mathbb{F}_3^{m \times n}$ of minimal weight in its coset with respect to the set of (I_2, J_1). Without loss of generality, we may assume that $I_2 = m$ and $J_1 = n$.

Observe that if there is a 1 in a row, there is no 2 in that particular row, otherwise, in the column of that 2 there is a 1 or a 2, since the sum of the entries of that column add to 1 mod 3. If we restrict our attention to the two columns and the two rows of the aforementioned 1 and 2 we have the following two possibilities:

$$\begin{pmatrix} 1 & 2 \\ * & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 2 \\ * & 2 \end{pmatrix}$$

where $*$ could be any symbol. Adding the codeword $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ we get a strictly lower weight with the same syndromes. Therefore, in every row, there are only 0's and 1's or only 0's and 2's. A similar result also holds for the columns.

Define:

$$\begin{aligned} I_{2,1} &= \{i \mid \text{row i has only 0's and 1's}\} \\ I_{2,2} &= \{i \mid \text{row i has only 0's and 2's}\} \\ I_{1,1} &= \{j \mid \text{row j has only 0's and 1's}\} \\ I_{1,2} &= \{j \mid \text{row j has only 0's and 2's}\}. \end{aligned}$$

Then, $I_2 = I_{2,1} \sqcup I_{2,2}$, and $J_1 = J_{1,1} \sqcup J_{1,2}$, where $\sqcup$ is the symbol for the disjoint union.

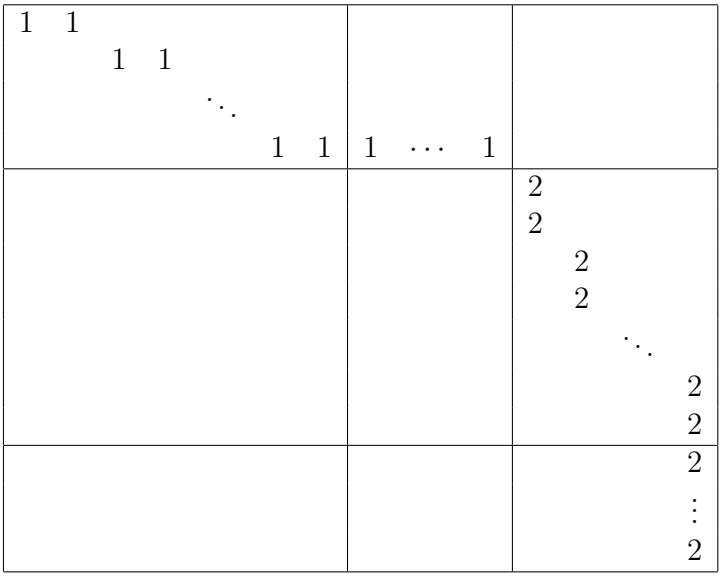
Observe that for all $i \in I_{2,1}$ there are $j, k \in J_{1,1}$ where $j < k$ with a 1 at (i, j) and (i, k). Hence the size of $\{j \mid x_{i,j} = 1\}$ is equal to 2 mod 3. Therefore $|J_{1,1}| \geq 2|I_{2,1}|$. Similarly, we have that $|J_{2,2}| \geq 2|I_{1,2}|$. Then,

$$\mathrm{wt}(\mathbf{e}) \geq |J_{11}| + |I_{2,2}|. \tag{1}$$

Furthermore, suppose that $j \in J_{1,1}$ and $i \in I_{2,1}$. Then there is at most one 1 in column j, since otherwise we have the following submatrix $\begin{pmatrix} 1 & 1 \\ 1 & * \end{pmatrix}$ and we can add $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ which gives an element in the same coset of strictly lower weight. So, for all $j \in J_{1,1}$, there is exactly one 1 in column j, and similarly for all $i \in I_{2,2}$, there is exactly one 2 in row i. Hence

$$\begin{aligned} \sum x_{i,j} &\equiv |J_{1,1}| + |I_{2,2}| \\ &\equiv 2|I_{2,1}| + |I_{2,2}| \\ &\equiv |J_{1,1}| + 2|J_{1,2}|. \end{aligned}$$

Therefore we have that $|J_{1,1}| \equiv |I_{2,1}|$ and $|J_{1,2}| \equiv |I_{2,2}|$, so $|J_{1,1}| - |I_{2,1}| \equiv 0 \bmod 3$ and $|I_{2,2}| - |J_{1,2}| \equiv 0 \bmod 3$, and the minimum weight of $\mathbf{e}$ stated in Equation 1 is attained, that is $\mathrm{wt}(\mathbf{e}) = |J_{11}| + |I_{2,2}|$, which is illustrated in the figure below:



Let $x = |I_{2,2}|$ and $y = |J_{1,1}|$. Recall that $|I_{2,1}| + |I_{2,2}| = m$, and $|J_{1,1}| + |J_{1,2}| = n$. Then $y = |J_{1,1}| \geq 2|I_{2,1}| = 2m - 2x$ and $y = |I_{2,2}| \geq 2|J_{1,2}| = 2m - 2y$.

Hence, we are looking for the values of x and y such that the weight of $\mathbf{e}$ is minimal. In other words, we are solving the following linear programming problem:

$$\min x + y$$

with constraints

$$\begin{aligned} 2m \leq y + 2x,\ 0 \leq y \leq n \\ 2n \leq x + 2y,\ 0 \leq x \leq m \\ x, y \in \mathbb{Z}. \end{aligned}$$

The minimal value of $x + y$ is obtained in a corner point of the feasible region. In our case, we have four corner points, that is

$$\begin{aligned}
&x = m, y = n, \text{ with weight } m + n \\
&x = m, y = \frac{1}{2}n, \text{ with weight } m + \frac{1}{2}n \\
&x = \frac{1}{2}m, y = n, \text{ with weight } \frac{1}{2}m + n \\
&x = \frac{4}{3}m - \frac{2}{3}n, y = \frac{4}{3}n - \frac{2}{3}m, \text{ with weight } \frac{2}{3}(m + n).
\end{aligned}$$

Since $m \leq 2n$, then it follows that $\frac{2}{3}(m+n) \leq n + \frac{1}{2}m$. Similarly, since $n \leq 2m$, $\frac{2}{3}(m+n) \leq \frac{1}{2}n + m$. Hence $\frac{2}{3}(m+n)$ is the minimum value for the problem.

Recall that $2|I_2| \equiv |J_1| \bmod 3$, hence $2m - n \equiv 0 \bmod 3$. Therefore $\frac{4}{3}m - \frac{2}{3}n = \frac{2}{3}(2m - n)$ is an integer. Similarly, since $m \equiv 2n \bmod 3$, also $\frac{4}{3}n - \frac{2}{3}m$ is an integer.

Similar proof also works in case $I_2 = \emptyset$, $J_1 = \emptyset$ and $2|I_1| \leq |J_2|$. ◇

Lemma 2.16 *Let* $\mathbf{x} \in \mathbb{F}_3^{m \times n}$. *Then*

$$\mathrm{wt}(\mathbf{x} + C_n \otimes C_m) = \begin{cases} \frac{1}{3}(|I_1| + |J_2|) + \frac{2}{3}(|I_2| + |J_1|) & \text{if } |I_1| < |J_1| \text{ and } |J_2| < |I_2| \\ \frac{1}{3}(|I_2| + |J_1|) + \frac{2}{3}(|I_1| + |J_2|) & \text{if } |J_1| < |I_1| \text{ and } |I_2| < |J_2|. \end{cases}$$

Proof.

Without loss of generality, we may assume that $I_0 = J_0 = \emptyset$, $|J_1| + |J_2| = n$, $|I_1| + |I_2| = m$ and $\mathbf{x}$ is rearranged such that the rows of I_1 are above the rows of I_2 and the columns of J_1 are before the columns of J_2.

Remark that $|J_1| + 2|J_2| \equiv |I_1| + 2|I_2|$ by Lemma 2.2.

Let $\mathbf{e}$ be a minimum weight element in $\mathbb{F}_3^{m \times n}$ having the same syndrome as $\mathbf{x}$. After a permutation of rows and columns $\mathbf{e}$ is partitioned as follows:

A	B	C
D	E	F
G	H	I

Where

$$\begin{aligned}
&A \text{ has size } |I_1| \times |I_1| \\
&B \text{ has size } |I_1| \times (|J_1| - |I_1|) \\
&C \text{ has size } |I_1| \times |J_2| \\
&D \text{ has size } |J_2| \times |I_1| \\
&E \text{ has size } |J_2| \times (|J_1| - |I_1|) \\
&F \text{ has size } |J_2| \times |J_2| \\
&G \text{ has size } (|I_2| - |J_2|) \times |I_1| \\
&H \text{ has size } (|I_2| - |J_2|) \times (|J_1| - |I_1|) \\
&I \text{ has size } (|I_2| - |J_2|) \times |J_2|
\end{aligned}$$

and every row of A are in I_1, and every row of D and G are in I_2, and also every column of A and B are in J_1, and every column of C are in J_2. The minimum weight of $\mathbf{e}$ can be attained by setting A to be the identity matrix, F to be the diagonal matrix with 2's on the diagonal, B, C, D, E, G, I are matrices with only zeros as entries, and H is filled using the same procedure as in Lemma 2.15.

Hence $\mathbf{e}$ has weight equal to

$$|I_1| + |J_2| + \frac{2}{3}\left[(|I_2| - |J_2|) + (|J_1| - |I_1|)\right].$$

Therefore $\mathrm{wt}(\mathbf{x} + C_m \otimes C_n) = \frac{1}{3}(|I_1| + |J_2|) + \frac{2}{3}(|I_2| + |J_1|)$.

The proof is similar in case $|J_1| < |I_1|$ and $|I_2| < |J_2|$.

◇

Proposition 2.17 *Let $\mathbf{x} \in \mathbb{F}_3^{m\times n}$. Let $I_i = I_i(\mathbf{x})$ and $J_i = J_i(\mathbf{x})$ for $i = 1, 2$. Then*

$$\mathrm{wt}(x+C_n\otimes C_m) = \begin{cases} |J_1| & \text{if } |I_1| = |J_2| = 0 \text{ and } 2|I_2| \le |J_1| \\ |J_2| & \text{if } |I_2| = |J_1| = 0 \text{ and } 2|I_1| \le |J_2| \\ |I_2| & \text{if } |I_1| = |J_2| = 0 \text{ and } 2|J_1| \le |I_2| \\ |I_1| & \text{if } |I_2| = |J_1| = 0 \text{ and } 2|J_2| \le |I_1| \\ |J_1| + |J_2| & \text{if } |I_1| \le |J_1| \text{ and } |I_2| \le |J_2| \\ |I_1| + |I_2| & \text{if } |J_1| \le |I_1| \text{ and } |J_2| \le |I_2| \\ \frac{2}{3}(|I_2| + |J_1|) & \text{if } I_1 = J_2 = \emptyset,\ |J_1| \le 2|I_2| \text{ and } |I_2| \le 2|J_1| \\ \frac{2}{3}(|I_1| + |J_2|) & \text{if } I_2 = J_1 = \emptyset,\ |J_2| \le 2|I_1| \text{ and } |I_1| \le 2|J_2| \\ \frac{1}{3}(|I_1| + |J_2|) + \frac{2}{3}(|I_2| + |J_1|) & \text{if } |I_1| < |J_1| \text{ and } |J_2| < |I_2| \\ \frac{1}{3}(|I_2| + |J_1|) + \frac{2}{3}(|I_1| + |J_2|) & \text{if } |J_1| < |I_1| \text{ and } |I_2| < |J_2|. \end{cases}$$

Proof. The above formula is a consequences of the Lemmas 2.13 - 2.15 ◇

Remark 2.18 *Since we can construct a coset leader of any $\mathbf{x}$ with minimum weight by Lemma 2.13 - 2.15, A complicated closed formula for the number of coset leader with certain weight could be derived, which we omit in this paper.*

References

[1] R. Jurrius and R. Pellikaan. The coset leader and list weight enumerator. *AMS, Contemporary Mathematics Series, vol. 632, pp. 229-252, 2015.* `http://www.win.tue.nl/~ruudp/paper/71.pdf`

[2] M. Munuera, Steganography from a coding theory point of view. *Algebraic Geometry Modeling in Information Theory (Edgar Martinez-Moro, ed.), vol. 8, World Scientific, New Jersey, 2013, pp. 83-128.*

[3] T. Helleseth and T. Klove The Newton radius of codes *IEEE Transactions on Information Theory, vol. 43, no. 6, pp. 1820-1831, Nov 1997.*

[4] MacWilliams, F J, and N J. A. Sloane *The Theory of Error-Correcting Codes* Amsterdam: North-Holland Pub. Co, 1996.

Secure comparison through simple bit operations

Thijs Veugen
TNO CWI
Unit ICT Cryptology Group
The Hague Amsterdam
thijs.veugen@tno.nl

Abstract

We present a solution to the well-known millionaires' problem, such that both parties only need to compute and communicate with bits. This avoids using complicated cryptographic libraries during implementation, and seriously reduces the computational and communication complexity compared to existing solutions. A trusted dealer is needed to generate and distribute sufficient random bits, which are needed during the execution of the protocol. The number of communication rounds is logarithmic in the number of input bits. Our protocol is secure in the semi-honest model.

1 Introduction

We present a solution to the well-known "millionaires' problem": party $\mathcal{A}$ has an integer a, party $\mathcal{B}$ has an integer b, and they would like to determine $a < b$, without revealing each others integer. This is usually solved by either homomorphic encryption, or secret-sharing. A typical solution for such a cryptographic protocol, called "secure comparison", is by Damgård, Geisler, and Krøigaard [3, 4, 7]. The disadvantage is when implementing, both parties need to have complex crypto functionality: generating keys, managing keys, encryption, decryption, computing with large numbers, etc.

We present a solution avoiding complicated hard- and software. The only operations that both parties need be able to do are straightforward computations with bits: multiplication and exclusive-or. The big advantage is that this makes the solution better scalable: simple clients can perform a secure comparison protocol, without having to install complicated crypto libraries, and without needing extra computing power. The challenge is to control the communication, more precisely the number of communication rounds, which usually take some time in practice. Furthermore, we need a third party to precompute a number of random bits (not related to the inputs) for the two clients.

2 Bitwise secret sharing

The basic idea is to binary secret-share every bit. Bit x is being split into two bits: x_A for party $\mathcal{A}$ and bit x_B for party $\mathcal{B}$, such that $x = x_A \oplus x_B$. We denote such a split secret bit x as $\langle x \rangle$. Adding (exclusive-or) of two secret bits is simple, since both parties can locally add their shares:

$$x \oplus y = (x_A \oplus y_A) \oplus (x_B \oplus y_B).$$

As with most cryptographic protocols, the challenge is the multiplication. To multiply two secret bits x and y to $z = x \cdot y$, we need three extra secret bits a, b, and c, such that $c = a \cdot b$. These three bits can be generated by an external party. The multiplication protocol [1] then looks as follows:

1. The parties $\mathcal{A}$ and $\mathcal{B}$ have two secret bits x and y, and a triplet of secret random bits a, b, and c, such that $c = a \cdot b$. All bits are secret-shared, for example, x is split into x_A and x_B, such that $x = x_A \oplus x_B$.

2. Both parties locally compute $\langle d \rangle = \langle x \rangle \oplus \langle a \rangle = \langle x \oplus a \rangle$ and $\langle e \rangle = \langle y \rangle \oplus \langle b \rangle = \langle y \oplus b \rangle$.

3. The parties reveal the secret bits d and e, by sending the privately held shares to each other.

4. Both parties locally compute $\langle z \rangle = \langle c \rangle \oplus (e \cdot \langle a \rangle) \oplus (d \cdot \langle b \rangle) \oplus (d \cdot e)$.

By working out the equation, one can see that the calculated secret bit z equals $x \cdot y$: $(d \oplus a)(e \oplus b) = (de) \oplus (db) \oplus (ae) \oplus (ab)$.

Both parties only need to perform a couple of bit operations on their privately held shares. The computation $e \cdot \langle a \rangle$ means that the parties multiply their local share with the known bit e, so $\mathcal{A}$ computes $e \cdot a_A$ and $\mathcal{B}$ copmutes $e \cdot a_B$, such that they jointly compute the new secret bit ea in a secure way.

Only step 3 requires communication, not more than two bits need to be sent to the other party.

3 Secure comparison

Having a solution to securely compute with binary values, i.e. multiplying and adding them, the next step is to build a secure comparison protocol with it [2]. The first step is to split the privately held integers a and b into ℓ bits, ℓ being the maximal bit length of the integers. Party $\mathcal{A}$ has bits $a_{\ell-1} \dots a_0$, denoting the integer a, and party $\mathcal{B}$ similarly has the bits $b_{\ell-1} \dots b_0$. These 2ℓ bits are easily turned into 2ℓ secret sharings, by setting the corresponding shares of the other party to 0.

Inspired by [3, 4], the secure comparison protocol then looks as follows:

1. For each i, $0 \leq i < \ell$, the parties $\mathcal{A}$ and $\mathcal{B}$ compute the secret bits $d_i = (a_i < b_i)$, by computing $\langle d_i \rangle = \langle b_i \rangle \cdot (1 \oplus \langle a_i \rangle)$. This takes ℓ secure multiplications, which can be executed in parallel, i.e. requiring one communication round.

2. The parties locally compute the secret bits $e_i = (a_i = b_i)$ by computing $\langle e_i \rangle = \langle a_i \rangle \oplus \langle b_i \rangle \oplus 1$ for each i, $1 \leq i < \ell$.

3. For each i, $0 \leq i < \ell$, the parties $\mathcal{A}$ and $\mathcal{B}$ compute the secret bit c_i by $\langle c_i \rangle = \langle d_i \rangle \prod_{j=i+1}^{\ell-1} \langle e_j \rangle$.

4. The parties locally compute the (secret-shared) output δ of the comparison: $\langle \delta \rangle = 1 \oplus \sum_{i=0}^{\ell-1} \langle c_i \rangle$.

We use the notation $d_i = (a_i < b_i)$ to denote the bit d_i that is 1, if $a_i < b_i$, and 0, otherwise. The same holds for $e_i = (a_i = b_i)$ and $\delta = (a < b)$. The correctness of the above protocol is shown similarly to [3].

The computational and communication (number of bits) complexity of the protocol above is very limited, compared to existing solutions. The reason is that we use secret sharing modulo two, which requires only computation with and communication of bits. The challenge is to reduce the $\ell - 1$ communication rounds for computing $\prod_{j=i+1}^{\ell-1} \langle e_j \rangle$.

4 Reducing the number of communication rounds

The challenge is, given secret bits $\langle e_i \rangle$, for $1 \leq i < \ell$, to compute the secret bits $\langle f_i \rangle$, such that $f_i = \prod_{j=i+1}^{\ell-1} e_j$, for $0 \leq i < \ell$. The straightforward approach is to perfrom the $\ell - 1$ multiplications sequentially, leading to $\ell - 1$ communication rounds.

A way to reduce the communication rounds from $\ell - 1$ to $\log_2(\ell - 1)$, is to compute the bits within a binary tree. This comes down to computing in round j all products with at most 2^j factors, $1 \leq j \leq \log_2(\ell - 1)$. In the first rounds one computes $e_1 e_2$, $e_3 e_4$, ... In the second round $e_1 e_2 e_3 e_4$, $e_5 e_6 e_7 e_8$, ..., is computed. And so on and so forth.

An example with $\ell = 9$ and three rounds:

1. Compute $f_8 = 1$, $f_7 = e_8$, $f_6 = e_7 \cdot e_8$, and further $e_1 \cdot e_2$, $e_3 \cdot e_4$, and $e_5 \cdot e_6$.
2. Compute $f_5 = e_6 \cdot f_6$, $f_4 = e_5 e_6 \cdot f_6$, and further $e_1 e_2 \cdot e_3 e_4$ and $e_2 \cdot e_3 e_4$.
3. Compute $f_3 = e_4 \cdot f_4$, $f_2 = e_3 e_4 \cdot f_4$, $f_1 = e_2 e_3 e_4 \cdot f_4$ and $f_0 = e_1 e_2 e_3 e_4 \cdot f_4$.

With this adjustment, the number of communication rounds seems manageable for practical applications. For example, with inputs being maximised to one billion, one needs $\ell = 30$ bits, which leads to $\lceil \log_2 \ell \rceil = 5$ rounds to compute all f_i. The total number of communication rounds then comes to $5 + 1 = 6$, where the exra round is needed to compute $c_i = d_i \cdot f_i$ (the computation of the d_i can be put into one of the five rounds).

5 Security

The described protocol is secure within the semi-honest security model [5]. This means that all participating parties are assumed to follow the rules of the protocol. This is sufficient for most practical applications. The trusted dealer, who generates the multiplication triplets, is not allowed to collude with either $\mathcal{A}$ or $\mathcal{B}$, otherwise the secret integer of the other party could be revealed. It is possible to avoid a trusted dealer and have the parties jointly generate the triplets. Although this approach is outside the scope of this paper, we suggest an efficient solution based on oblivious transfers, as described in [6], to jointly generate the triplets.

6 Conclusions

We described a very efficient solution for the millionaires' problem, which avoids the use of complicated cryptographic hard- and software, and only requires computations and communication of bits. This could be easily used in practice whenever a trusted dealer is available, without leading to serious performance loss (compared to the non-secure version). The number of communication rounds is logarithmic in the number of input bits.

Acknowledgement

The research activities that have led to this paper were partly funded by PPS-surcharge for Research and Innovation of the Dutch ministry of Economic Affairs.

References

[1] D. Beaver. One-time tables for two-party computation. In Computing and Combinatorics, pages 361–370. Springer, 1998.

[2] Martine De Cock and Rafael Dowsley and Caleb Horst and Raj Katti and Anderson C. A. Nascimento and Stacey C. Newman and Wing-Sea Poon, Efficient and Private Scoring of Decision Trees, Support Vector Machines and Logistic Regression Models based on Pre-Computation, Cryptology ePrint Archive: Report 2016/736, `https://eprint.iacr.org/2016/736.pdf`, 2016.

[3] I. Damgård and M. Geisler and M. Krøigaard, Homomorphic encryption and secure comparison, Journal of applied cryptology, vol. 1, no. 1, pp. 22–31, 2008.

[4] I. Damgård and M. Geisler and M. Krøigaard, correction to efficient and secure comparison for on-line auctions Journal of applied cryptology, vol. 1, no. 4, pp. 323–324, 2009.

[5] Oded Goldreich, Foundations of Cryptography: Basic Applications, vol. 2, Cambridge University Press, 2004.

[6] Vladimir Kolesnikov and Ranjit Kumaresan, Improved OT extension for transferring short secrets, Crypto 2013, pp. 54–70, 2013.

[7] Thijs Veugen, Improving the DGK comparison protocol, IEEE Workshop on Information Forensics and Security, December 2012.

Rate-Distributed Spatial Filtering Based Noise Reduction in Wireless Acoustic Sensor Networks

Jie Zhang, Richard Heusdens, Richard C. Hendriks
Faculty of Electrical Engineering, Mathematics and Computer Science,
Delft University of Technology, 2628 CD Delft, The Netherlands
{j.zhang-7, r.heusdens, r.c.hendriks}@tudelft.nl

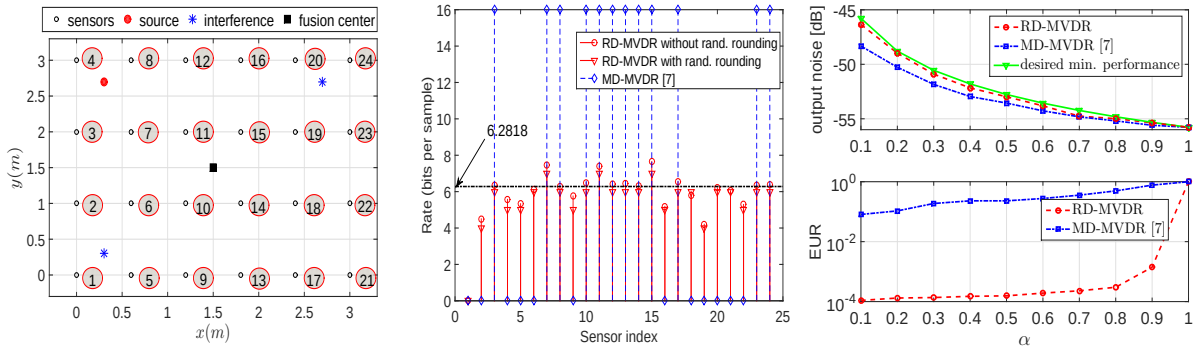
Abstract

Nowadays, wireless acoustic sensor networks (WASNs) have attracted an increasing amount of interest. Compared to conventional microphone arrays with a fixed configuration, WASNs are better scalable without strict array size limitations. In WASNs, each sensor node is usually battery powered having a limited energy budget. It is important to take the energy consumption into account in the design of algorithms to prolong the network lifetime.

To reduce the energy usage, there are two techniques that can be employed: sensor selection and rate allocation. For sensor selection, the most informative subset of sensors is chosen by maximizing a performance criterion while constraining the cardinality of the selected subset, or by minimizing the cardinality while constraining the performance. In this way, the number of sensors contained in the selected subset can be much smaller than the total set of sensors, resulting in a sparse selection, but each selected sensor quantizes data at full rate. Therefore, much less data need to be processed at a fusion center (FC), i.e., sensor selection can efficiently save the energy usage in terms of data processing. Compared to sensor selection, rate allocation allows for a more smooth operating curve as sensors are not selected to only operate at full rate or zero rate (when not selected), but at any possible rate. For rate allocation, the idea is to allocate higher rates to the more informative sensors while lower or zero rates are allocated to less informative sensors. Since the transmission cost between a sensor and the FC is exponential in the rate, rate allocation can save more energy in terms of data transmission, even though all the sensors could still be involved. Hence, the difference between sensor selection and rate allocation problems lies in binary versus more smooth decisions.

In this work, we only consider the energy usage for data transmission and neglect the energy usage for other processes. The wireless transmission power is regarded as a function of the distance between sensor nodes and the FC and the rate (i.e., bit per sample). We minimize the battery usage due to transmission, while constraining the noise reduction performance. This results in an efficient rate allocation strategy, which depends on the underlying signal statistics, as well as the distance from sensors to the FC. Under the utilization of a minimum variance distortionless response (MVDR) beamformer, the problem is derived as a semi-definite program. Furthermore, we show that rate allocation is more general than sensor selection, and sensor selection can be seen as a special case of the presented rate-allocation solution, e.g., the best microphone subset can be determined by thresholding the rates.

Fig. 1(a) shows the experimental setup consisting of 24 microphones in a 2D room with dimensions 3×3 m. One target speech source and two interfering sources are present, and the FC is placed at the center of the room. Fig. 1(b) shows an example of the rate distributions of the sensor selection method (MD-MVDR) and the proposed rate allocation method (RD-MVDR). We can see that MD-MVDR activates less sensors, each at the maximum rate of 16 bits per sample; RD-MVDR has more active sensors, each at much lower rate. Fig. 1(c) shows the output noise power and the energy usage ratio. It can be seen that both RD-MVDR and MD-MVDR can satisfy the desired minimum performance, but the RD-MVDR method is more efficient in energy usage.



(a) (b) (c)

Figure 1: (a) experimental setup, (b) rate allocation example for a single target source (the maximum rate is fixed to 16 bits) and (c) output noise power and energy usage ratio (EUR, which is defined as the ratio between the energy usage of the RD-MVDR/MD-MVDR method and the maximum energy usage when all the sensors are involved and each at the maximum rate) in terms of α.