Proceedings of the 34rd WIC Symposium on Information Theory in the Benelux

and

The 3rd Joint WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux

Leuven, Belgium
May 30–31, 2013

Co-organized by imec and KU Leuven
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
15. 1994 Louvain-la-Neuve, Belgium ISBN 90-71048-10-1
20. 1999 Haasrode, Belgium ISBN 90-71048-14-4
23. 2002 Louvain-la-Neuve, Belgium ISBN 90-71048-16-0

Proceedings

Edited by Sofie Pollin, Liesbet Van der Perre, Annemie Stas, Leuven Universiteit Twente Bibliotheek
The 34rd Symposium on Information Theory in the Benelux and The 3rd JointWIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux have been organized by

Circuits and Systems for ICT group, Smart Systems and Energy Technology division, imec
Networked Systems group, ESAT-TELEMIC division, KU Leuven

on behalf of the

Werkgemeenschap voor Informatie- en Communicatietheorie, the IEEE
Benelux Information Theory Chapter and the IEEE Benelux Signal Processing Chapter.

Organizing committee
Liesbet Van der Perre (imec & KU Leuven)
Sofie Pollin (KU Leuven & imec)
Annemie Stas (imec)

The organizing committee gratefully acknowledges the financial support of the Gauss Foundation for the “Best Student Paper Award”, of the IEEE Benelux Information Theory Chapter, and the WIC for the “Best Student Presentation Award”. 
Preface

The Werkgemeenschap voor Informatie- en Communicatie-theorie (WIC) has organized the annual Symposium on Information Theory in the Benelux (SITB) since 1980. This year’s symposium, the 43rd in the series, takes place in Leuven, Belgium. For the third time, it is organized jointly with the IEEE Benelux Signal Processing Chapter. The symposium is co-organized by imec and KU Leuven, by wireless and information systems enthusiasts from the circuits and systems for ICT and networked systems groups. These proceedings contain the papers which are presented during the symposium. We are grateful to the authors for sharing their latest research with us.

This year we are extremely fortunate to have two interesting keynote speeches, from abroad and from within imec: Prof. Ove Edfors (Lund University, Sweden) and Dr. Andy Lambrechts (imec, Belgium).

We gratefully acknowledge the sponsorship provided by the Gauss Foundation (presenting the Best Student Paper Award) and the IEEE Benelux Chapter on Information Theory. Also, the WIC sponsored the Best Student Presentation Award.

We also express our sincere thanks to Ms. Annemie Stas for her great help in the organization and hosting of the symposium.

We hope that this symposium offers a good opportunity to exchange knowledge and improve personal contacts among the participants, and results in novel scientific, cultural and personal discoveries.

Leuven, May 2013,
Liesbet Van der Perre and Sofie Pollin
(Symposium Organizers)
Program of the WIC symposium 2013

Location: imec, Caf -1A

Thursday, May 30, 2013

09:00 – 10:00  Registration/coffee

10:00 – 10:10  Opening by Liesbet Van der Perre

10:10 – 10:50  Invited presentation of Professor Ove Edfors,
Professor at Lund University, 221 00 Lund, Sweden:
Massive MIMO - Performance of low-complex linear precoding

Spatial multiplexing using massive MIMO has been shown to have very promising properties, including low-complex linear precoding, large increases in spectral efficiency, and several orders of magnitude lower transmit powers - as compared to today's access schemes. These things have, however, been shown for theoretical channels with i.i.d. complex Gaussian coefficients. From a practical point of view, it is therefore important to evaluate massive MIMO in more realistic scenarios, with real massive-MIMO channels. We pursue this by analyzing measurement data from several measurement campaigns in the 2.6 GHz frequency range, using different antenna array structures with 128 antenna elements. A comparison between i.i.d. and measured massive-MIMO channels is done, with the objective of finding out how well we can expect low-complex linear precoding to work in real massive-MIMO environments.

10:50 – 11:10  Coffee break

11:10 – 12:40  Cryptography

T. Veugen, H. Stokking: Secure Processing Offload in Recombining Media Segments for Mobile Access
(TNO)

(Istanbul Technical University, KU Leuven)

Q. Wang, V. Rijmen, Deniz Toz, K. Varici: Study of the AES-like Super Boxed in LED and PHOTON
(KU Leuven)

(Delft University of Technology)

12:40 – 14:00  Lunch

14:00 – 15:30  Signal Processing for Communication

A. Martin, L. Vandendorpe, J. Louveaux: Precoder Optimization for DSP with High Crosstalk
(Université Catholique de Louvain)

M. Bauduin, T. Deleu, F. Duport, P. De Doncker, S. Massar, F. Horlin: Equalization of the non-linear 60 GHz channel: Comparison of reservoir computing to traditional approach
(Université Libre de Bruxelles)
E.J.G. Janssen, H. Habibi, Wu Yan, J.W.M. Bergmans, P.G.M. Baltus: A system Study on Nonlinear Interference Suppressor for Local Interferences (Eindhoven University of Technology)

C. Li, M. Li, S. Pollin, M. Verhelst, L. Van der Perre: Towards perfectly rejecting one dominating harmonic interference in SDR receiver (imec, KU Leuven)

Poster session and refreshment drinks

T. Vermeulen, S. Pollin: Digital Self-Interference Cancellation for Full Duplex Communication on USRP (KU Leuven)


H. Habibi, E.J.G. Janssen, Wu Yan, P.G.M. Baltus, J.W.M. Bergmans: Closed-loop adaptation of Non-linear Interference Suppressor for Local Interferences (Eindhoven University of Technology)


J.-F. Determe, a. Bourdoux, F. Horlin: An OMP algorithm with memory for spectrum sensing using the Modulated Wideband Converter (Université Libre de Bruxelles)

Z. Jin, T. Wang, J.-B. Wei, J. Louveaux, L. Vandendorpe: A Low-Complexity Algorithm for sum Rate Maximization in Multi-Cell Opportunistic DF relay aided OFDMA downlink systems (Université Catholique de Louvain)


L. De Meyer, B. Bilgin, B. Preneel: Extended Analysis of DES S-boxes (KU Leuven)

A. Hendrikse: Signal Processing in hearing aids (GN ReSound)

X. Wang, J.-P. Linnartz, T. Tjalkens: Optimization by simulated annealing: Intelligent light control using an array of dimmable LEDs (Eindhoven University of Technology)

17:00 – 18:00 General Assembly of the Werkgemeenschap voor Informatie- en Communicatietheorie

18:30 – 19:30 Visit Museum M (Meeting point: in front of the Museum M)

19:30 Dinner at Museum M
Friday, May 31, 2013

09:00 – 10:30  **Capacity and Coding**

L. Tolhuizen: *Coding for known erasure values at the decoder*  
(Philips Research Eindhoven)

A. Tsiatmas, C.P.M.J. Baggen, F.M.J. Willems: *Information Transmission using Illumination Systems*  
(Eindhoven University of Technology)

(Delft University of Technology)

(Eindhoven University of Technology, University of Twente, MESA Institute for Nanotechnology)

10:30 – 11:00  Coffee break

11:00 – 12:30  **Poster session 2**

C.J.A. Jansen: *On the Orders of Binary Matrix Pairs*  
(Compumatica, The Netherlands)

L. Tolhuizen: *The error probability after erasure decoding is monotonically non-decreasing in the erasure probability*  
(Philips Research Eindhoven)

V. Lazov, S. Pollin, G. Vandenbosch: *Indoor Propagation and Radiation Measurements for Wireless Communication Systems*  
(KU Leuven)

K. Papagiannopoulos, G. Alpar, W. Lueks: *Designated Attribute Proofs with the Camenish-Lysyanskaya Signature*  
(Radboud University Nijmegen, TNO)

E. Marin, D. Singelee, S. Pollin: *Security Analysis of An implantable Cardioverter Defibrillator*  
(KU Leuven)

J. Rodriguez, M. Mercuri, P. Karsmakers, P.J. Soh, P. Leroux, D. Schreurs: *Automatic Fall Detector Based on Sliding Window Principle*  
(KU Leuven, Thomas More Kempen)

C. van Dam, L.J. spreeuwers, R.N.J. Veldhuis: *Model-free 3D Shape Reconstruction from Video Sequences*  
(University of Twente)

J. de Groot, J.-P. Linnartz: *Secrecy rate versus verification performance in biometric authentication schemes*  
(Eindhoven University of Technology)

12:30 – 14:00  Lunch
Sensing and Feature Detection

Y. Peng, L. Spreeuwers, B. Gokberk, R. Veldhuis: Comparison of Super-Resolution Benefits for Face Recognition on Downsampled Images and Realistic Low-Resolution Data
(University of Twente)

(Université Libre de Bruxelles)

A. Jalalirad, T. Tjalkens, J.-P. Linnartz: Selecting feature-based models
(Eindhoven University of Technology)

(Delft University of Technology)

15:10 – 15:50  Invited presentation of Dr. Andy Lambrechts
imec, Leuven, Belgium:

Hyperspectral imaging: from lab to industry

Although the potential of hyperspectral imaging has been demonstrated for many applications, using laboratory setups in research environments, its adoption by industry has so far been limited due to the lack of high speed, low cost and compact hyperspectral cameras. To bridge the gap between research and industry, we have developed a novel hyperspectral sensor that integrates the spectral filters directly on top of a standard CMOS imager. The spectral filters can be matched with the application and have been demonstrated in both line-scan and snapshot implementations that address different application areas. The result is a compact and fast hyperspectral camera made with low-cost CMOS process technology, enabling hyperspectral imaging in industry, medical and consumer applications.

15:50 – 16:00  Best student paper award ceremony

16:00  Closing by Liesbet Van der Perre
Secure Processing Offload in Recombining Media Segments for Mobile Access

Thijs Veugen, Hans Stokking
TNO, Technical Sciences
P.O. Box 5050, 2600 GB, Delft, The Netherlands
{thijs.veugen, hans.stokking}@tno.nl
Thijs Veugen is also affiliated with the Information Security and Privacy Lab of Delft University of Technology

Abstract
We develop an architecture for a federation of home gateways. We distinguish inner and outer circle friends and describe a mechanism for social-aware backup of content and sharing with friends. To facilitate remote access, content can be segmented and redundantly stored at gateways of friends. We develop innovative solutions that allow secure and redundant storage of segments with inner and outer circle friends. They enable decoding of segments in the encrypted domain without the need of sharing a decryption key with friends, which relieves the computational effort of mobile devices accessing the content. We combine Vandermonde-Reed-Solomon codes with homomorphic encryption to three different solutions and describe their differences in complexity.

I. INTRODUCTION
Nowadays, people create, share and consume content in a multitude of ways. Using their mobile devices, people take photos and shoot videos, and post them directly online to share with their friends, family and often with larger groups in their social network.

The FP7 FIGARO project focuses, among other things, on secure distributed backup, sharing and remote access to content, using people’s social network. To enable this, the project has created an architecture of federated home gateways of the users. This federation allows a user’s home gateway to make backups at the gateways of friends and family. These backups can be ‘social’ backups, to be used for the purposes of sharing the content with friends and enabling access to content while on the move [1][2].

If video content is of high quality, the bandwidth demands for streaming this content will be (much) higher than the average uplink of a single home gateway. Therefore, for the purpose of remote access, the backup of the (video) content is segmented into small parts. This allows for streaming of high-quality (HD) video, using the uplink bandwidth of multiple friends’ gateways at the same time. Each gateway can then stream certain segments, to the extent the uplink bandwidth permits.

This paper focuses on secure remote access. We distinguish between two groups of friends, the so-called ‘inner circle’ and ‘outer circle’ friends. Inner circle friends are friends whom you completely trust, while outer circle friends are from the larger group in your social network, whom you do not trust completely.

We use Vandermonde-Reed-Solomon codes [4] to ensure the back-up can be retrieved as long as a limited number of home gateways is online. This is a common technique in storage technology also known as Redundant Array of Independent Disks (RAID). To avoid leakage of content towards outer circle friends we use homomorphic encryption [5]. The homomorphic property enables outer circle friends to decode segments in the encrypted domain, thereby reducing the workload of the downloading client to decryption only. The application of additively homomorphic encryption, but also of multiplicatively homomorphic encryption is demonstrated.

To further reduce the workload of the downloading (mobile) client, a solution based on symmetric encryption is described which accelerates the client’s decryption effort. As with all our solutions, to reduce security risks we don’t allow inner circle friends to learn the client’s private key.

In Section II we describe our new architecture for social-aware backup and sharing of content. In Section III the security measures are described and analyzed, and compared with existing solutions. We end with the conclusions in Section IV.
FIGARO proposes an evolvable future Internet architecture based on gateway-oriented federation of residential networks. The residential gateway has a central role in the FIGARO vision of the future Internet. It interconnects the residential network with the Internet and is responsible for aggregating a multitude of devices and services within the residential network. In FIGARO, residential gateways undertake the federator role, internally as well as externally. Figure 1 shows residential networks connected at the edge of the Internet and illustrates a simplified view including the two types of residential network federations. The upper part illustrates external federation interconnecting multiple gateways to form a cooperative overlay across residential networks. This federation enables further collaboration to offer added value in terms of, for example, access and sharing of content, storage and network capacity. The right-part of the figure shows the internal federation within a residential network.

A. **Social-aware backup and sharing**

Backup provides durability through duplication, while caching raises availability of content. The mechanisms of these functions can be combined to establish network efficiency when both doing backups and sharing content in a group of close friends. E.g. when backing up your family pictures, you can create a backup at your families gateways and at the same time share the content with your family. The content is then transmitted over the network only once, while serving the two purposes of both backup and sharing, including mobile access (see next subsection). This backup is comparable to work presented in [3]. Social backup is different in the sense that it is based on close relationships, e.g. family and good friends. We envision having offline contact with these relations, and agreeing on the use of the others’ bandwidth and storage. This usage does not have to be reciprocal. A main insight here is that people you trust and know, are willing to help you, and this circumvents the idea of freeriding.

![Figure 1: Overview of the FIGARO environment](image)

<table>
<thead>
<tr>
<th>Client device</th>
<th>Federation manager</th>
<th>GWx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Management</td>
<td>External Federation Control &amp; Management</td>
<td>Content Management</td>
</tr>
<tr>
<td>External Federation Control &amp; Management</td>
<td>Lookup Service</td>
<td>External Federation Control &amp; Management</td>
</tr>
</tbody>
</table>

1. Request available friends’ GWs
   (Reply: Available GWs)

2. Request backup

3. Check authentication
   (Reply: Backup accepted)

4. Transfer Backup Segment(s)

5. Notify Content Location

![Figure 2: Social-aware backup performed by a client device to a GWx of a friend](image)
The social-aware backup protocol in the FIGARO environment is shown in Figure 2. In this figure, a client device is shown making a backup of some (piece of) content at a gateway $x$ (GW$x$) from some friend. In a first step, the client device requests a list of available friends’ gateways from the Lookup Service. The Lookup Service is the part of the centrally located Federation Manager which amongst other things keeps track of all available gateways and of the users of those gateways. After receiving a list of available gateways and making a selection, the client device requests the backup from GW$x$, belonging to a friend. After GW$x$ has authenticated the client device as belonging to a friend, the content is optionally encrypted (for explanation, see below). Then, the backup itself is performed. This backup is performed in segments to enable remote (mobile) access, see next subsection.

Whether the content is encrypted, depends on the role of the friend in the user’s social network. We have divided the social graph of a user in two parts: inner circle friends and outer circle friends, as shown in Figure 3. Inner circle friends are the friends or family with whom you have a very strong connection, i.e. there is a high level of trust between you and them. Outer circle friends are the friends with whom you do connect online, but with whom the connection is less strong. For your inner circle friends, encryption of content is not performed, while for outer circle friends, it is.

After backing up the content, the user can now share this content as well. If the content is shared with a user from a location at which a backup is placed, only an authorization step is necessary, as the content itself is already distributed. This is shown in Figure 4. The client device first requests the friend’s GW location from the Lookup Service, and then sets the access rights for the other user in the AA module at the Federation Manager. These access rights are used in the mobile access scenario in the next subsection. Finally, a notification of the shared content is sent to the other GW, and optionally a key for access to the content.

![Diagram showing the process of sharing content](image-url)
Although the encryption itself does not need to be special for this type of backup and sharing, there are some new requirements to be met here. If different content is shared with different users, different keys need to be used in each occasion. While some content may be shared with a certain user, other content may not, while both content items can be backed up at that user’s location. This requires different keys for different content items.

B. Mobile access

To enable remote access to content, either by the owner of the content or by users with whom the content is shared, content can be segmented. Reason for this is the limited upload network bandwidth most connections have. E.g. streaming of a high-quality video may require a 4 mbit/s connection, while many upload speeds are limited to e.g. 512 kbit/s or 1 mbit/s. By segmenting the content in small segments, and distributing these segments across various locations, even high-quality video becomes remotely accessible for streaming. This can thus be seen as a form of caching, included for free (i.e. without making additional copies in the network) in the backup and sharing scenario. For granting other users such access when sharing content, the FIGARO Federation Manager keeps track of these authorizations in the central AA function.

Figure 5 shows this mobile access scenario. The mobile device first has to discover the location of the backed-up content segments, and can then request these segments. Gateways containing these segments first have to check if the requesting device has authorization, and can then deliver the segments. If the users of the mobile device and the gateway(s) containing the segments are inner circle friends, then the segments will be delivered unencrypted. This saves the mobile device processing and thus precious battery life. If the users are outer circle friends, segments will be encrypted and it is up to the mobile device to decrypt the segments before having access to the content.

III. Security measures

In order to realize both the “social-aware backup and sharing” as well as the “remote access” scenario in a secure way we need a couple of security measures to adequately protect content from outer circle and other unauthorized users, and assure its availability. A frequently used encryption system in such environments which offers different keys for each content is known as Convergent encryption [9], where the key is more or less a cryptographic hash of the content. Although this could be used in the “social-aware backup and sharing” scenario as described in Subsection II-A, the segmentation of content required in “remote access” asks for a more sophisticated solution.

To reduce content leakage we want to avoid storing a decryption key for each content (movie, document, etc.) at inner circle locations. The solution we propose uses Vandermonde-Reed-Solomon codes and homomorphic encryption. We first explain these two main concepts and then show how to combine these.

A. Vandermonde-Reed-Solomon codes

The first concept, also known as RAID, enables storing the content in a distributed way such that whenever a certain amount of users is available, the content can be retrieved. This measure guards the availability of content. More precisely, we store $n$ chunks of the content, represented by some integer $X$, at various locations such that $k$ chunks, $k < n$, are sufficient to restore the content.
Similar to [4], the key ingredient here is a $n \times k$ Vandermonde like matrix $M$ that has the property that any $k$ rows are linearly independent. This can be achieved [6] by defining the elements of $M$ for each $j, j = 1, \ldots, k$, as:

$$M_{ij} = \begin{cases} \delta_{ij} & \text{when } i = 1, \ldots, k; \\ j^{k-i} & \text{when } i = k+1, \ldots, n. \end{cases}$$

Here $\delta$ denotes the Kronecker delta. By using $M$, the content can be distributed and reconstructed as follows:

1. Divide content $X$ into $k$ integer valued segments $X_1, \ldots, X_k$, each of size $|X|/k$.
2. Compute the $n$ content chunks $(Y_1, \ldots, Y_n)^T = M^T (X_1, \ldots, X_k)^T$ and store each chunk at a different location.
3. Suppose $k$ chunks are available, represented by a subset $S$ of $\{1, \ldots, n\}$ of size $k$. Let $M_S$ be the matrix consisting of these $k$ rows of $M$, and let $Y_S$ be the vector consisting of these $k$ chunks.
4. Since the $k$ rows of $M_S$ are linearly independent, this matrix can be inverted and the $k$ segments can be reconstructed by $(X_1, \ldots, X_k)^T = M_S^{-1} * Y_S^T$.
5. The content $X$ can be downloaded and combined from the $k$ segments $X_1, \ldots, X_k$.

Because the segments and the matrix elements are integers, the chunks $Y_1, \ldots, Y_n$ will also be integers. However, because the matrix elements can be quite large, the size of the chunks might grow leading to storage size disadvantages. To overcome this problem, the computations in steps 2 and 4 could be performed in a finite field of size $N$ such that $N > |X|/k$ [6]. This condition assures correct reconstruction of the segments.

B. Homomorphic encryption

With respect to the confidentiality requirement, our solution should enable inner circle friends to retrieve the content but should avoid outer circle friends learning the content. For this purpose we introduce the second concept known as homomorphic encryption which is a form of encryption that allows a limited number of operations on encrypted data.

We distinguish between additively homomorphic encryption and multiplicatively homomorphic encryption and give examples for both systems. A well-known encryption system that is additively homomorphic is Paillier [7]. In Paillier, the encryption of $x, 0 \leq x < N$, is $[x] = g^x * r^N \mod N^2$, where $g$ is a generator, $r$ a fresh random value, and $N$ a large RSA-like number consisting of the product of two large primes. Paillier is additively homomorphic because $[x] * [y] = [x + y] \mod N^2$.

A well-known multiplicatively homomorphic encryption system is (unpadded) RSA [8]. In RSA, an encryption of $x, 0 \leq x < N$, is $[x] = x^e \mod N$ for some (public) integer $e$ and number $N$ that consists of the product of two large primes. RSA is multiplicatively homomorphic because $[x] * [y] = [x * y] \mod N$. The encryption $[x]$ is decrypted by raising it to the secret power $d$, where $d$ is the multiplicative inverse of $e \mod \phi(N)$.

In both encryption systems, one has to know the prime factors of $N$ to be able to decrypt.

C. Secure and redundant storage with inner and outer circle friends

The concepts of redundant storage and homomorphic encryption can be nicely combined to a secure system that fulfills the requirements of both scenarios. We describe how content could be stored redundantly and securely, and retrieved by mobile devices, using additively homomorphic encryption. To reduce the workload of mobile devices retrieving the content from the network, we would like to have RAID decoding done by the inner and outer circle friends.

1. The content owner generates an instantiation of Paillier and broadcasts the public key to his inner circle friends.
2. To store content $X$, he divides content $X$ into $k$ integer valued segments $X_1, \ldots, X_k$, each of size $|X|/k$.
3. He computes the $n$ chunks $(Y_1, \ldots, Y_n)^T = M^T (X_1, \ldots, X_k)^T$ and stores each chunk at a different inner circle friend.
4. Inner circle friends are allowed to store copies of their chunks $Y_i$ at outer circle friends, but only after encrypting it with the content owner’s public key: $[Y_i]$.
5. Suppose $k$ encrypted chunks are available from outer circle friends, represented by a subset $S$ of $\{1, \ldots, n\}$ of size $k$. Let $M_S$ be the matrix consisting of these $k$ rows of $M$, and let $Y_S$ be the vector consisting of these $k$ chunks.

$$Y_S^T = M_S^{-1} * X_i^T$$
6. Since the $k$ rows of $M_S$ are linearly independent, this matrix can be inverted and the $k$ segments can be reconstructed by $(X_1, \ldots, X_k)^\top = M_S^{-1} \cdot Y_S^\top$. Because the chunks are now encrypted, reconstruction now has to be performed by outer circle friends in the encrypted domain:

$$[X_i] = \prod_{j \in s} [Y_j]^a_{ij} \mod N^2$$

where the numbers $a_{ij}$ are the elements of the matrix $M_S^{-1}$.

7. The content owner downloads the encrypted segments $[X_i]$ from the outer circle friends and decrypts them with his private key.

8. The content owner downloads and combines $X$ from the $k$ segments $X_1, \ldots, X_k$ and enjoys the content.

Note that the matrix $M$ should be available to inner circle friends to be able to enjoy the shared content. This could be arranged by the Federation Management component as described in the previous section. Also outer circle friends that need to reconstruct the (encrypted) segments need to know the elements of the matrix $M_S^{-1}$. This could be achieved through inner circle friends or directly from the Federation Management component.

A disadvantage of Paillier is that encryption blows up the information size by a factor two, namely from $N$ to $N^2$. Furthermore, decryption of segments by the mobile device of the content owner will be costly. This can be solved by a subtle change in encryption: instead of encrypting his chunk $Y_i$ with the content owner’s public key in step 4, the inner circle friend could encrypt his chunk with a self-generated symmetric key, e.g. by using Convergent encryption [9], and encrypt this symmetric key with the content owner’s public key. This reduces storage size and simplifies the decryption by the mobile device, but on the other hand prevents the network from doing the RAID decoding for the user.

Alternatively, RSA could be used with avoids blowing up the size by encryption. However, because it is not additively but multiplicatively homomorphic, the encoding process has to be modified.

The chunks $Y_j$ should be computed through exponentiation instead of multiplication (as in step 3):

$$Y_j = \prod_{i=1, \ldots, k} X_i^{M_{ji}} \mod N$$

The reconstruction of segments in the encrypted domain (as in step 6) however is unmodified:

$$[X_i] = \prod_{j \in s} [Y_j]^a_{ij} \mod N$$

where the numbers $a_{ij}$ are the elements of the matrix $M_S^{-1}$. Note that the relations in the plain domain are different: $X_i = \sum_{j \in s} a_{ij} \cdot Y_j \mod N$ with Paillier, but $X_i = \prod_{j \in s} Y_j^{a_{ij}} \mod N$ with RSA. The computation of a matrix inverse requires the computation of multiplicative inverses. In Paillier these inverses have to be computed in the field $\mathbb{Z}_{N^2}^*$, but in RSA these are done in the field $\mathbb{Z}_{\phi(N)}$ which raises a problem because $\phi(N)$ cannot be computed without knowing the factorization of $N$. However, by representing the numbers $a_{ij}$ as rational numbers and choosing integer $c$ large enough such that all numbers $c \cdot a_{ij}$ are integers, outer circle friends will be able to reconstruct the segments $[X_i]^c$. The content owner is able to decrypt (in step 7) these segments to $X_i$ by raising them to the power $d \cdot c^{-1} \mod \phi(N)$.

D. Related work

A solution known from the sensor node domain [10][11] is to use symmetric encryption and secretly share the symmetric key with adjacent nodes, or in our domain with the inner circle friends. The disadvantage of this solution is that any $k$ inner circle friends are able to retrieve the decryption key. An outsider somehow getting access to a sufficient number of key shares would also be able to retrieve the decryption key and thus obtain the content through (untrusted) outer circle friends. A second disadvantage of this solution is that content always has to be decrypted before usage, even by inner circle friends, which causes an undesirable computational effort.

To the authors’ knowledge, our solution is unique in the sense that it enables RAID like decoding in the encrypted domain without the need of sharing the decryption key.

E. Complexities

In Subsection III-C we have described three different solutions that use RAID but avoid sharing of the decryption key by inner circle users. The computational and communication complexities of these three solutions are compared, both for the (mobile) client and for the network.
In the first solution, the client generates a Paillier key pair, and distributes the public key among inner circle friends. The public key is used by inner circle friends to encrypt their content chunks $Y_i$ before storing it at outer circle friends. RAID decoding is done by the network in the encrypted domain.

The second solution is similar but uses RSA instead of Paillier which avoids data expansion by encryption but requires a different coding scheme. The main differences with the first solution are:

1. Different encryption and decryption algorithms (RSA vs. Paillier).
2. The size of an encryption is $N^2$ in Paillier but only $N$ in RSA.
3. The computation of chunks $(Y_1, \ldots, Y_n)$ from segments $(X_1, \ldots, X_k)$ requires only multiplications with Paillier but exponentiations with RSA.
4. The computation of segments $(X_1, \ldots, X_k)$ from chunks $Y_S$ requires exponentiations to the (rational) powers $a_{ij}$ with Paillier but (integer) powers $c \ast a_{ij}$ with RSA. For small values of $k$ and $n$ the integer powers are smaller than the rational powers (of size $N$), but since a similar trick can be used with Paillier, this difference is negligible. However, because of Paillier’s larger size of encryptions, each multiplication (and exponentiation) will require a larger effort than with RSA.
5. Because we use unpadded RSA, this crypto system is less secure than Paillier, which is semantically secure [7].

The third solution uses any public key encryption system (not necessarily homomorphic), and the public key is distributed among inner circle friends. Inner circle friends use Convergent encryption to encrypt their content chunks $Y_i$ before storing it at outer circle friends. The symmetric key is also transmitted to outer circle friends after encryption with the public key. This solution increases the speed of encryption and decryption, but doesn’t allow RAID decoding in the encrypted domain. The main differences with the first solution are:

1. Convergent encryption uses fast symmetric encryption and decryption algorithms. Public key encryption is only used for encrypting (small) symmetric keys.
2. Convergent encryption doesn’t lead to a blowup in encryption size, but requires additional transmissions of the (encrypted) symmetric keys.
3. The computation of segments $(X_1, \ldots, X_k)$ from chunks $Y_S$ requires exponentiations with Paillier, but is done in the plain domain with Convergent encryption and thus requires only multiplications.
4. The computation of segments $(X_1, \ldots, X_k)$ from chunks $Y_S$ is done by the network with Paillier, but has to be performed by the client with Convergent encryption.

To conclude, the second, RSA-based solution seems to be preferable to the first, Paillier-based solution from a communication and storage complexity point of view. The disadvantages of the second solution being the higher computational complexity of the encoding scheme (to compute the chunks from the segments) and the weaker crypto system.

From a communication and storage complexity point of view, the third, Convergent encryption based solution is only slightly worse than the second one because of the symmetric keys that have to be additionally stored. From a computational point of view, the big advantage of the third solution is the fast encryption and decryption of chunks, but a substantial amount of work is transferred to the (mobile) client which might outweigh this advantage.

IV. CONCLUSIONS

In the FIGARO project we developed an architecture for a federation of home gateways. We distinguished inner and outer circle friends and described a mechanism for social-aware backup of content and sharing with friends. To facilitate remote access, content can be segmented and redundantly stored at gateways of friends.

We developed innovative solutions that allow secure and redundant storage of segments with inner and outer circle friends. They enable decoding of segments in the encrypted domain without the need of sharing a decryption key with friends, which relieves the computational effort of mobile devices accessing the content. We combined Vandermonde-Reed-Solomon codes with homomorphic encryption to three different solutions and described their differences in complexity.
ACKNOWLEDGEMENT

This work has been financed by European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2009-5-258378 (FIGARO project).

REFERENCES


Implementation of a Lightweight Trusted Platform Module

Mehmet Akif Ozkan$^{1,3}$, Anthony Van Herreweghe$^2$, Ingrid Verbauwhede$^2$, S. Berna Ors$^{1,4}$

$^1$: Faculty of Electrical and Electronics Engineering, Istanbul Technical University  
{akif.ozkan, siddika.ors}@itu.edu.tr
$^2$: ESAT/COSIC-iMinds, KU Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium  
{anthony.vanherrewege, ingrid.verbauwhede}@esat.kuleuven.be
$^3$: Erasmus Student at KU Leuven University
$^4$: ITU Faculty of Electrical Engineering Room 2318, Maslak, Istanbul

Abstract: Today’s computing platforms are becoming more and more mobile and networked, while their tasks get increasingly critical. Therefore, the need to verify and identify that a local or remote computing platform behaves as expected has become an important challenge. Software and hardware attestation protocols have been proposed to solve this problem in the past few years. While many vulnerabilities and attacks have been discovered against the proposed software based solutions, hardware based solutions are too costly for lightweight embedded devices. Recently, lightweight solutions requiring minor hardware changes have been proposed for the low-end embedded devices. One of the state of the art approach is SMART (Secure and Minimal Architecture for Root of Trust), which memory accesses are controlled by looking at the program counter (PC), offered by El Defrawy et al. [1].

In this work, the SMART protocol is implemented using a soft-core 8051 microprocessor. The Dalton core [2] is modified and used for the implementation of the 8051 microprocessor on a Xilinx Virtex4 FPGA. SHA3 algorithm (KECCAK[r=40, c=160] function and 224 bit output length is chosen) is used as the hash function. SDCC is used as compiler. Cost, compactness and design effort of the SMART scheme are examined for this restrictive microprocessor. Although it is necessary to add new ROM partitions for the attestation software and key, it is seen that the SMART don’t need too many changes to the existing architecture. Moreover, the maximum operating frequency of 11.6 MHz is decreased by a negligible amount (<0.01%) in the final design.

1. Introduction

Remote attestation is a two party protocol used to gain assurance that a local or remote computing platform behaves as expected. The term challenger is used for the verifier that checks the internal state of the device called prover over a network.

While some protocols can only offer static attestation, some of them provide dynamic attestation by giving the information of the state of a platform at run time. Software based solutions are proposed to be used in resource constrained platforms [3-7]. However, many vulnerabilities and attacks have been discovered against them [8-10]. Current remote software attestation methods rely on too restrictive assumptions for many realistic applications and expect no collusion which means that the prover should be linked directly with the verifier without using a network. Existing hardware based approaches based on adding a chip that is called Trusted Platform Module (TPM) [11], which cannot even be compromized by the owner, are proposed by the Trusted Computing Group (TCG). Although one can make provable statements about the platform configuration and protect keys from multiple users by using TPMs and secure coprocessors, these approaches are focused on general-purpose hardware while embedded systems are neglected.

Recently, lightweight solutions requiring minor hardware changes have been proposed for the low-end embedded devices [12-13,1]. One approach that uses physical unclonable
function (PUF) hardware, which is a physical one way function, with a combination of a software based solution is proposed by Schulz et al. [12]. The self-protecting module structure of Strackx et al. [13] and SMART (Secure and Minimal Architecture for Root of Trust) by El Defrawy et al. [1] are other promising methods that control memory accesses by looking at the program counter (PC) and establish dynamic attestation.

In this work, SMART architecture is investigated and implemented for the 8051 microprocessor. Dalton soft core of the 8051 is used and final design is implemented in the Virtex-4 FPGA. The Security analysis and design goals of the system are discussed at the chapter 2. Then the design of SMART for a general architecture is presented at chapter 3. Next, implementation details of the architecture are shown for the 8051 microarchitecture.

2. Security Considerations and Design Elements

The Remote attestation protocol that uses SMART is given in Figure 1. It is assumed that a secure algorithm, e.g. HMAC, calculates a cryptographic checksum of the prover’s state using a key and nonce when the challenger requests. The key is used to prevent fake replies and the random nonce is used for freshness. The challenger checks the reply of the prover to provide assurance.

Certain assumptions have been made about the adversary in the SMART system [1]. The adversary can have a complete control over the system before and after attestation execution. He can control and change the software, code and data of the prover and he can control the communication channel any time. However, the adversary can’t perform hardware attacks on the prover, e.g. changing Smart code in ROM, learning the key using side channel attacks. It is also assumed that the prover and the verifier shared a secret key, which can be loaded at production time or later in a secure way. The SMART protocol is developed to provide attestation where the prover is a low-end embedded device that has been modified to have the following characteristics [14]:

- The device should have single memory space. There should be no separation between “kernel” and “user” memory.
- The device should have single thread of execution, with the exception of interrupts. It means that the device shouldn’t offer Direct Memory Access (DMA) or that DMA can be securely disabled during attestation.
- The device should have the ability to disable interrupts and force a region of code to execute atomically.
- The device should have a read-only memory (ROM).
- The device should have the ability to clean up memory upon device reset.
- A hardware-based control mechanism to prevent unauthorized access to certain memory locations.

From the definition, remote attestation must satisfy the following two properties to protect following two attack scenarios [14]:

- **Property of Attestation 1**: The attest can only compute the valid checksum since only it has access of the correct key.
  - **Attack scenario 1**: The adversary simulates the attestation.
- **Property of Attestation 2**: Different checksum results should be calculated for different states. \( \text{Attest} (s_1,k) \neq \text{Attest} (s_2,k) \) must be for different states \( s_1 \neq s_2 \).
  - **Attack scenario 2**: The adversary can corrupt attestation of the state for an incorrect calculation of checksum.
In order to protect the system from these two attack scenarios three security objectives are summarized by El Defrawy et al. [1]:

- **Prover Authentication**: The verifier authenticates the prover
- **External verification**: The verifier is guaranteed that the prover has the expected content for the questioned memory segment \([a, b]\).
- **Guaranteed execution**: The verifier is guaranteed that the prover executed the requested code location.

The main goal of SMART is to realize these security objectives with minimal hardware changes. This is done with a piece of code and architectural changes to the platform that provides a set of features, which are given as necessary and sufficient by Francillon et al. [14].

- **Exclusive access to key**: Attestation must have exclusive access to key. A privileged mode can be used if the processor supports multiple privileged modes and separation of memory for each process. Since low-end embedded devices generally don’t have this feature, it is suggested by El Defrawy et al. [1] that a piece of hardware can be added to enforce that the key is only accessible from the attestation region by checking the program counter (PC) and the address bus.
- **No Leaks**: Attest mustn’t leak any information related to the key except the result. Any intermediate values should be erased after the attestation to prevent leaking.
- **Immutability**: The attestation code must be immutable to prevent the adversary from changing the code to learn the key. In order to protect the attestation code, it is placed in ROM [14].
- **Uninterruptibility**: The adversary can change the place of malware during attestation to hide it from detection. Attestation must run from the start to end without interrupted. For single threaded execution, interrupts should be disabled at the beginning.
- **Invocation from start**: The adversary can make attestation interruptable by invoking the code from anywhere or he can check the sanity checks on input parameters. Therefore the attestation code must be forced to run from the start.

Both uninterruptibility and invocation from start properties together provide atomic execution that mostly protects from scenario 2 attacks. Exclusive access, no leaks and immutability are necessary properties to prevent scenario 1 attacks. The set of features are realized with the SMART architecture using four main building blocks for the prover:
• **Attestation Read-Only Memory**: Attestation code in ROM. Key only accessible from here.
• **Secure Key Storage**: Memory region inside CPU to store key.
• **MCU Access Controls**: Custom hardware to control exclusive access to key and atomic execution.
• **Reset And Memory Erasure**: Secure reset mechanism with reliable and secure memory erasure

### 3. Attestation Architecture

SMART is a challenge response attestation architecture that uses special hardware features of the prover. The memory architecture of the platform is shown in the figure 2. Both key and SMART ROM code (RC) are placed in different memories in order to prevent any information leakage about the key ignoring the physical attacks. The ROM that the attestation algorithm resides is placed in the address space of the microarchitecture. The key is stored in a special hardware controlled memory that is placed at the address space. The memory controller hardware gives exclusive access to RC code and checks if the RC code is started from the first address and executed atomically. This is managed by checking the PC and data address. Internal reset is given to the system by the memory controller hardware for any violation of these principals. Giving the assurance that the checksum is calculated by the RC code yields external execution.

![Figure 2: Memory Architecture Overview of SMART](image)

The attestation algorithm of the RC is shown in algorithm 1. HMAC or public key digital signature algorithms can be used for the RC code. The attestation starts after the verifier sends the input parameters of the algorithm, which are shown in algorithm 1, to the prover. Then, the RC computes a checksum using nonce over the memory region \([a, b]\) in the prover’s memory. The random number must have sufficient bit length to keep the result unique enough for each request. Calculating the checksum correctly gives fresh authentication of the prover and this yields prover authentication to the verifier. After the calculation of the checksum, the RC passes control to the function at the x address if the \(x_{flag}\) is set. Otherwise it returns to the user program after interrupts are restored. If the flag is set and x is set in the attested region \((x \in [a, b])\), the verifier can gain the assurance that the program will execute the expected code. This yields guaranteed execution.
Algorithm 1: SMART Attestation Algorithm in ROM (RC Code)

input: a, b; start/end addresses for attestation
  x, address to jump to after attestation
  x\text{flag}, decision flag to jump x address after attestation
  n, nonce number sent by the verifier
  out, output address to store checksum
  in, (optional) input parameter for the function at the x address

output: C, checksum

body:

1: Disable interrupts
2: Check validity of the out address and stack pointer
3: Check validity of the x address
4: Compute the checksum C, using HMAC of range \([a, b]\) and nonce
5: Clean all temporary variables from memory
6: Write the checksum to the output address
7: if \(x\text{flag} == 1\) then // Check if x flag is set
8: Jump \((x, \text{in})\); // Jump to the function at the address x with its input
9: else
10: Restore interrupts;
11: end

At the start of the algorithm the output address, stack pointer and the jump address \(x\) must be checked to avoid attacks [1,14]. For example, a corruption is possible if the stack pointer is addressed to IO or the output address to where checksum will be written points in the stack [1]. The jump address \(x\) cannot be set in the RC region [14]. A denial-of-service-attack is possible by addressing the last instruction as \(x\) address. In this way, an infinite loop can be created although the attested code is verified. Moreover, SMART can be invoked partially by addressing \(x\) to somewhere in the RC code. This violation can’t be understood by the system because the program counter never leaves the RC code.

An interrupt can be set to occur for RC execution interval. The PC will never leave from SMART ROM code if the service routine of that interrupt is addressed in RC. To avoid this, interrupts must be disabled at the first step of the RC as seen in algorithm 1. However disabling interrupts is not necessary for microarchitectures whose interrupt service routines are fixed and cannot be addressed in the RC code. Because leaving the RC while execution will cause a violation of atomic execution and the controller memory trigger the reset as designed.

SMART can’t reserve data memory, use global registers or the heap to avoid relying on trusted data [1]. So the RC code must work on the stack.

4. Implementation

The Intel 8051 is an 8-bit microcontroller that has 16-bit address bus and 8-bit data bus. It has two types of memory that are program memory (ROM) and data memory (RAM). The ROM is used to save the program to be executed and the RAM is used to store data and intermediate results during the operation. Depending on the model of the 8051 family, it has a few kB of ROM and 128 or 256 bytes of RAM. It is also possible to add external memory chip with capacity of 256 kB. The Dalton core [2] that has 256 byte of internal RAM and 128 Kb external RAM is used in this paper. The 256 byte internal RAM has two partitions that are 128 byte of low internal RAM and 128 bytes of high internal
RAM. Special Function Registers, which are used to monitor and control the microcontroller—such as hardware counters, timers—are built in the high internal RAM partition. It has four register banks with eight registers at the low internal RAM. While the low internal RAM is addressable for both direct and indirect access, higher internal RAM and external RAM can only be accessed with indirect addressing.

4.1. Software Design for SMART

The memory area of the RC code is very important for architectures like the 8051, which has very limited resources. A lightweight implementation of the KECCAK algorithm (SHA3) [17] is chosen to design keyed hash function (MAC) for the attestation Code (RC).

The security strength of the HMAC function is mainly defined from the security of its inner hash. The KECCAK function doesn’t have the length-extension weakness like SHA1 and SHA2. Therefore MAC computation can be performed by simply prepending the message with the key instead of the HMAC nested construction [17]. The RC can concatenate the message (attested region) with the key and nonce and then calculate the hash of it to find checksum, whenever the attestation is requested. Instead of this, a faster way is used for the implementation in this work. The KECCAK hash result of the key is stored instead of the key. This hashed key is initialized at the start of the KECCAK algorithm to calculate the hash result of the message and nonce when attestation is required. In this way, hash calculation of the key isn’t done in every time.

The KECCAK algorithm takes a message and calculates its hash with using a 5*5 state array. The bit width of the state b is determined by the sum of two parameters that are capacity c and bitrate r. They are set to 160 and 40 respectively to design lightweight algorithm that uses 8-bit word for its state array. The half of the capacity gives the bit level security strength level of the algorithm and for this design it is $2^{80}$. The algorithm calculates the hash result of the bitrate length message, which is 40 bit for this design, at a time.

The KECCAK algorithm mainly has two partitions. First of it is the KECCAK-f sponge function. This function takes the state array as input and gives as output after it makes certain mathematical calculations on the state array for at least 12 iterations. The iteration number, which is 18 for this design, is related with the bit width of the state. The second part of the KECCAK algorithm has three main phases that are padding, absorbing and squeezing. After the message is padded and partitioned to bitrate width pieces, each partition and the previous result of the state are XORed and given as input to the KECCAK-f sponge function at the absorbing phase. Finally, if it is necessary a few more iterations are done at the squeezing phase. In each of these iterations, Keccak-f function is executed and the first bitrate width of its output are concatenated to produce the desired output length.

224 bit output width, which is the smallest of the SHA3 standard, is chosen for the checksum. 10 byte length is chosen for the key and the nonce of RC design since it is considered as sufficient.

SDCC is used as the compiler for the 8051 microarchitecture in this work. Since using the internal RAM with direct accessing gives the fastest code, SDCC assigns every variable as global and addresses them to the internal RAM when the program code is compiled for the fastest result. However, RC can only work on the stack and all intermediate variables must be cleaned after the execution as the design rule of SMART. Reading and writing variables from the stack is only possible with indirect access. This decreases the speed and increases the size of the algorithm by a factor two. The state array is written to the stack and other variables are written to registers.
The sponge function is written in ASM for a fast program and the top-level control is written with C to gain flexibility. As a result, 16 registers and 25 bytes of the stack are used from internal memory. The time of the attestation algorithm is as 1844 instruction-cycle/byte. Note that the term cycle is used for clock cycle, 1 instruction-cycle is 12-clock cycle. The necessary code storage for the RC is 1263 bytes. No other 8051 implementation of the KECCAK[r=40, c=160] algorithm is found at the time this work was written. However, this implementation is considered good by comparing with the AVR implementation of KECCAK[200] [18] and the 8051 implementation of KECCAK[1600] [19] which are 2411 and 1648 instruction cycle/byte respectively.

4.2. Hardware Design for SMART

The RC memory is put in program address space, key memory is put in data address space as seen in figure 2. Additional hardware that controls memory accesses to the key and guarantees atomic execution is designed. The exclusive access to key memory hardware is given at the figure 3.

![Figure 3: The Schematic view of the hardware that controls the exclusive access to key][1]

The total memory of the system isn’t changed. Therefore the total hardware cost of SMART is considered as the sum of the additionally required hardware, key memory and RC code memory. The designed additional hardware uses 1-Flip Flop (FF) and 28 4-input Look Up Tables (LUT). The FF is used to design the atomic execution circuit, which allows the program to only enter at the first address of RC and exit from the last. Because there is no pipeline in 8051 microarchitecture, the additional hardware to control memory accesses increases the critical path by 156 ns, which reduces the maximum default operating frequency of 11.6 MHz by a negligible amount (<0.01%).

5. Conclusion

In this work, an implementation of SMART for the 8051 is presented. The SMART architecture as one of the state of the art solutions is examined and implemented for the 8051 microarchitecture. The Dalton core is modified and implemented in a Virtex-4 FPGA. A lightweight version of KECCAK (SHA3) algorithm is used as attestation algorithm. It is
shown that while the necessary hardware changes to the platform are negligible, the cost mostly comes from the software that is used for the attestation code.

References


Study of the AES-like Super Boxes in LED and PHOTON

Qingju Wang\textsuperscript{1,2}, Vincent Rijmen\textsuperscript{\dagger}, Deniz Toz\textsuperscript{\dagger} and Kerem Varıcı\textsuperscript{\dagger,*}

\textsuperscript{1} KU Leuven Department of Electronical Engineering ESAT SCD-COSIC and iMinds
Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, Belgium
\textsuperscript{2} Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, 200240, China.
{qingju.wang, vincent.rijmen, deniz.toz, kerem.varici}@esat.kuleuven.be

Abstract. In this paper we investigate the choice of S-box and MDS-matrix in AES-like super box structures. Before giving the theory, we first study the super boxes of the LED block cipher and the PHOTON hash function and compute the distribution of all two-round characteristics when there is a key addition in the round function. Then, we compare these results with a fixed key (i.e. using the given constant values) in the round function. We observe that the results are consistent with the theory (following the approach of plateau characteristics) except for some extreme cases. Next, we discuss the effect of the S-box size for a fixed size super box. We show that the number of right pairs increases if the size of S-box decreases. Moreover, we study the effect of MDS-matrix with the previous setting by using a circulant matrix and the special MDS-matrix used in the design of LED.

1 Introduction

Over the past decades, especially with the introduction of modern computers and the rapid growth in communication, the field of cryptography became essential even in daily life. Being the most widely used cryptographic primitive, block ciphers play an important role in cryptography: Besides their traditional role encrypting data, they are also used as building blocks for hash functions and random number generators.

The block cipher Rijndael was designed by Daemen and Rijmen and it was selected as Advanced Encryption Standard (AES) in 2000. Since its introduction, due to its simplicity, the ease of analysis and efficiency, it has influenced the design of other cryptographic algorithms. Nowadays, there are many algorithms based on the strategy of AES or directly using the components of AES \cite{1–3,11}. A remarkable progress in the cryptanalysis of these algorithms has been made. Some recent results \cite{3,16,17} show the importance of the super box structure which is the smallest block in the algorithm which contains a linear and a non-linear element. In the AES-like designs, two rounds of the algorithm can be replaced by super boxes. Therefore, results on super boxes also represent the results for the two rounds of the algorithm.

Moreover, plateau characteristics were introduced and used in the super box analysis of AES in \cite{9}. This work focuses on the fixed-key probability of characteristics, denoted by DP\textsuperscript{[k]} (which is also an important criterion for block ciphers) and shows that the dependency on the value of the key is very structured for AES.

Our Contribution. We extend the work of \cite{9} to the block cipher LED \cite{13} and hash function PHOTON \cite{12}. These two algorithms are specifically designed for hardware purposes and use AES-like primitives. Many new lightweight cryptographic designs, including LED and PHOTON, use four-bit S-boxes rather than eight-bit to reduce the

\*This work was sponsored by the Research Fund K.U.Leuven, OT/08/027
necessary area requirements in hardware. Moreover, the MDS-matrix is also changed from a circulant matrix to a special MDS-matrix which is more hardware friendly.

In this work, we investigate the effect of S-box size and the choice of MDS-matrix on AES-like structures. For this purpose, we adopt the idea of previous work on plateau characteristics to study the super box structure and focus LED and PHOTON that use the same $4 \times 4$-bit S-box used in PRESENT [6] as case study. In order to observe the effect of the MDS-matrix, we investigate what happens if we replace it by a circulant MDS-matrix. We provide results for two rounds of LED, PHOTON-80/20/16 and PHOTON-128/16/16. Based on our results, we are able to compare super boxes of same size using different S-boxes and MDS matrices.

Outline. This paper is organized as follows. In Section 2, we first briefly describe the block cipher LED and the hash function PHOTON. Then we give the required definitions for the rest of the paper. In Section 3, we further study the plateau characteristics and then give an algorithm to compute height of a given characteristic in a super box. Using this algorithm, we compute the distribution (i.e., number of characteristics) for the LED and PHOTON super boxes, both for the keyed version and for the given round constants. Furthermore, we discuss the effect of the S-box and the MDS-matrix on the distribution in Section 4. Finally, we conclude this paper and summarize our findings in Section 5.

2 Definitions and Background

In this section, we first briefly describe the LED block cipher and PHOTON hash function. Then, we give some necessary definitions that ease the understanding of the paper.

2.1 LED block cipher

LED [13] is a conservative lightweight block cipher whose design can be seen as a special case of the generalized Even-Mansour construction [10]. LED accepts a 64-bit plaintext $P$, represented by a $4 \times 4$ array, and a 64-bit (or 128-bit) user key as inputs, and is composed of 8 (or 12) STEP functions preceded by a key addition.

The STEP function is an AES-like design composed of four rounds. Each round is a combination of Constant addition, S-boxes, ShiftRows, and (a variant of) MixColumns. LED uses the PRESENT [6] S-box. In the MixColumns operation, each column vector is multiplied by a matrix. The round constants for the second column are obtained from a linear shift register while the round constants for the remaining three columns do not change. Since we will focus only on the STEP function of LED, the key schedule is not mentioned in this paper. For more details, we refer to the specification of LED [13].

2.2 PHOTON hash function

PHOTON is a lightweight hash function family. The sponge construction [4, 7] which is based on a $t$-bit permutation $P$, with capacity $c$ and rate $r$, is used in the design of PHOTON. In the initialization phase the message is first padded and divided into $r$-bit message blocks. Then, these message blocks are xored to the internal state interleaved

---

\[1\] Remark: All ShiftRows and MixColumns used in this paper are not those of AES but transformation with similar properties.
with the permutation operation. When all message blocks are used, rate \( r' \) bits of internal state are extracted after each permutation call until \( n \) bits of output (the desired hash size) is reached. The former step is named as the absorbing phase and the latter is the squeezing phase.

Each member of the PHOTON family is named as PHOTON-\( n/r/r' \). The permutation used in PHOTON is the iteration of an AES-like round function for 12 times. In each round, a Constant addition is followed by SubCells, ShiftRows and MixColumns\(^1\). Depending on the hash size, the PRESENT [6] or the AES [8] S-boxes are used in the SubCells operation. An \( m \times m \) MDS-matrix is used in the MixColumns operation where \( m \) changes with the hash size. In our analysis, we focused on PHOTON-80/20/16 and PHOTON-128/16/16. We refer to the specification of PHOTON [12] for the rest of the details.

## 2.3 Planar differentials and plateau characteristics

The first definition is related with the use of S-boxes and linear transformations in the round function. Previously, the combination of these two structures is defined as a *super box*.

**Definition 1 (Super box [9]).** A super box maps an array of \( m \) elements \( a_i \) to an array \( e \) of \( m \) elements \( e_i \). Each of the elements has size \( n \). A super box takes a key \( k \) of size \( m \times n = n_b \). It consists of the sequence of four transformations (layers): Substitution, Mixing, Round Key Addition, Substitution.

Including LED and PHOTON, all AES-like structures can be described as a super box structure. In contrast to AES, the key addition is replaced by the constant addition in LED and PHOTON. The structure is depicted in Figure 1 for \( m = 4 \). Four parallel instances of the super box structure give an equivalent description for two rounds of LED whereas we need five and six parallel instances of the super box structure for PHOTON-80/20/16 and PHOTON-128/16/16, respectively. However, analysis of one of the super boxes is sufficient to deduce the results for two rounds of the given algorithms.

![Fig. 1. Super box representation for \( m = 4 \)](image)

**Differential Cryptanalysis** [5] is one of the most widely used techniques in cryptography and also many new techniques have been introduced later based on this idea. It investigates the relationship between the input and output differences (generally xor) through a target function. Checking the resistance of an algorithm (block ciphers, stream ciphers, hash functions, etc.) against differential cryptanalysis starts with analyzing the smallest non-linear component (mostly S-boxes).

Denote a *differential characteristic* \( Q \) as a sequence of differences \( Q = (\Delta_0, \Delta_1, \cdots, \Delta_m) \) through various stages of the encryption. Then, a *differential* [14] can be shown as
(\Delta_0, \Delta_m) where \(\Delta_0\) is the input difference and \(\Delta_m\) is the output difference. The differential probability (DP) of a characteristic or a differential over a map is the fraction of pairs that satisfies the given characteristic or differential and denoted as DP\((Q)\) or DP\((\Delta_0, \Delta_m)\), respectively.

Then, over a keyed map, DP\([k](Q)\) and DP\([k](\Delta_0, \Delta_m)\) are used to define differential probabilities of a characteristic and a differential for each key value \(k\). The expected differential probability (EDP) is the average of the differential probability over all keys. The weight of a differential or a characteristic is given as minus the binary logarithm of their EDP.

**Planar Differentials.** Let \(\gamma\) be a map and let \(F(a,b)\), \(G(a,b)\) be the sets that contain the input values, respectively the output values, for the right pairs of the differential \((a, b)\). i.e., \(F(a,b) = \{x | \gamma(x) + \gamma(x + a) = b\}\) and \(G(a,b) = \gamma(F(a,b))\). A differential \((a, b)\) is called a planar differential, if \(F(a,b)\) and \(G(a,b)\) form affine subspaces [9]. In that case, we can write:

\[
F(a,b) = p + U(a,b) \\
G(a,b) = q + V(a,b),
\]

where \(U(a,b)\) and \(V(a,b)\) are uniquely defined vector spaces, \(p\) any element in \(F(a,b)\) and \(q\) any element in \(G(a,b)\). Obviously, the sets \(F(a,b), G(a,b), U(a,b), V(a,b)\) have the same cardinality. Note that, if a differential \((a, b)\) has exactly two or four right pairs, then it is always planar [9].

**Plateau characteristics [9]** are a special type of characteristics whose probability for each value \(k\) of the key, DP\([k](Q)\), depends on the key and can have only two values. For a fraction

\[
2^{n_b - (\text{weight}(Q) + \text{height}(Q))}
\]

of the keys \(DP[k](Q) = 2^{\text{height}(Q) - n_b}\) \(1\)

and for all other keys it is zero. Here, \(\text{height}(Q)\) is an integer number.

### 3 The plateau characteristics in super boxes

**Two-Round Plateau Characteristic Theorem** [9, Theorem 1] states that a characteristic \(Q = (a, b, c)\) over a map consisting of two steps with a key addition in between, in which the differentials \((a, b)\) and \((b, c)\) are planar, is a plateau characteristic with \(\text{height}(Q) = \dim(V(a,b) \cap U(b,c))\).

In this section, we first give an algorithm to compute the height of a given characteristic based on this fact. We then use this algorithm not only to compute the full characteristic distribution of a given cipher but also to check how many of these characteristics are actually satisfied for the given constants of this super box.

#### 3.1 Algorithm to find the characteristic distribution for super boxes

A differential characteristic through the AES-like (including LED and PHOTON) super boxes consists of a sequence of four differences: the input difference \(a\), the difference after the first substitution \(b\), the difference after the mixing step (denoted as \(M\)) which is equal to the difference after the round (key) constant addition \(d\), and the output
difference after the second substitution $e$. These characteristics are denoted by $Q = (a, b, d, e)$.

For LED and PHOTON, both the sub-characteristics $(a, b)$ and $(d, e)$ are planar, therefore using the approach in Two-Round Plateau Characteristic Theorem, the height of the right pair $\text{height}(Q)$ can be determined by $\text{dim}(M(V(a,b)) \cap U(d,e))$, where $M(V) = \{M(v)|v \in V\}$ and vectors $v$ are $n_n$ bits vectors. The aim of our algorithm is to build the matrix $B$ containing the basis vectors of $M(V(a,b))$ and $U(d,e)$, and compute $\text{height}(Q) = \text{dim}(M(V(a,b))) + \text{dim}(U(d,e)) - \text{dim}(M(V(a,b)) + U(d,e))$ which is equal to the number of linearly independent rows of $B$.

The first step of our algorithm is to determine $V(a,b)$ and $U(d,e)$. Since, the super box is a set of $m$ parallel maps, $V(a,b)$ and $U(d,e)$ can be written as:

$$V(a,b) = V(a_1,b_1) \times V(a_2,b_2) \times \cdots \times V(a_m,b_m)$$
$$U(d,e) = U(d_1,e_1) \times U(d_2,e_2) \times \cdots \times U(d_m,e_m)$$

by using the Lemma 4 in [9] with “×” denoting the direct product [15]. If $|G(a_i,b_i)| > 0$, we are interested in the output values of the right pairs.

- If $|G(a_i,b_i)| = 2$, then the right pairs have input values in the set $q_i + \{0, b_i\}$ for some $q_i$ in $G(a_i,b_i)$, the basis vector for $V(a_i,b_i)$ being $b_i$, thus $V(a_i,b_i) = \langle b_i \rangle$.
- If $|G(a_i,b_i)| = 4$, then the right pairs have input values in the set $q_i' + \{0, b_i, \beta_i, b_i + \beta_i\}$ for some $q_i'$ in $G(a_i,b_i)$. Then, we take the primary difference $b_i$ and secondary difference $\beta_i$ as the basis vectors, therefore the input vector can be represented as $V(a_i,b_i) = \langle b_i, \beta_i \rangle$.
- If $(a_i, b_i) = (0, 0)$, then $G(a_i,b_i)$ covers the whole space and $V(a_i,b_i) = \langle w_0, w_1, \cdots, w_{n-1} \rangle$, where $w_j$ is a coordinate vector (i.e. a vector with 1 at position $j$ and zero at all other positions) and $V(a_i,b_i)$ is the standard basis.

If $|F_{(d_i,e_i)}| > 0$, we can use the input values of the right pairs for each parallel map to compute the height by Algorithm 1. The number of dependent rows in matrix $B$ gives $\text{dim}(M(V(a,b)) \cap U(d,e))$ which is equal to the height.

We would like to emphasize that, when there is key addition between the maps $(a, b)$ and $(d, e)$, the right pairs exists for a fraction of the keys and their values differ depending on the key. On the other hand, if the constant addition is used instead of the key addition operation in the cipher (as in including LED and PHOTON), it is not always guaranteed to have right pairs.

When constant addition $\text{const}$ is used, the right pairs can be determined by intersecting the affine spaces $F_{(a,b)} \cap (G_{(b,c)} + k)$. This can be efficiently done by adding one additional row, to be more precise $M(p) + q + \text{const}$, to the matrix $B$ and checking if this extension increases the rank.

Algorithm 1 calls the following subroutines: $\text{Add}(v)$ adds the vector $v$ as a new row to the matrix $B$. $\text{RowReduce}$ is the Gaussian Elimination and $\text{RowCount}$ gives the number of nonzero rows of a matrix. $\text{HEIGHT}$ computes the height of the characteristic when there is key addition, and $\text{CHECK}$ controls whether the characteristic has a solutions for the given constant.
Algorithm 1 Algorithm to compute the height of a given plateau characteristic

**Input:** Characteristic \( Q = (a, b, d, e) \) with \( \text{EDP}(Q) > 0 \) and constant \( \text{const} \)

**Output:** \( \text{height}(Q) \)

1: procedure PRECOMPUTE
2: for \( i = 1 \rightarrow m \) do
3: Compute \( G(a_i, b_i) = p_i + \langle b_i, \beta_i \rangle \) and \( F(d_i, e_i) = q_i + \langle d_i, \delta_i \rangle \)
4: end for
5: end procedure

6: procedure GENERATE_B
7: //at the input of Mixing
8: for \( i = 0 \rightarrow m \) do
9: if \( b_i = 0 \) then
10: for \( j = 0 \rightarrow n \) do
11: Add(\( M(w_{4i+j}) \))
12: end for
13: else if \( b_i > 0 \) then
14: Add(\( M(b_i) \))
15: if \( |V(a_i, b_i)| = 4 \) then
16: Add(\( M(\beta_i) \))
17: end if
18: end if
19: end for
20: //at the output of Mixing
21: for \( i = 0 \rightarrow m \) do
22: if \( d_i = 0 \) then
23: for \( j = 0 \rightarrow n \) do
24: Add(\( w_{4i+j} \))
25: end for
26: else if \( d_i > 0 \) then
27: Add(\( d_i \))
28: if \( |U(d_i, e_i)| = 4 \) then
29: Add(\( \delta_i \))
30: end if
31: end if
32: end for
33: end procedure

34: procedure HEIGHT
35: \( B' = \text{RowReduce}(B) \)
36: return \( \text{height}(Q) = \text{RowCount}(B) - \text{RowCount}(B') \)
37: end procedure

38: procedure CHECK
39: \( B'' = \text{RowReduce}(B') \)
40: if \( \text{RowCount}(B'') = \text{RowCount}(B') \) then
41: return true
42: else
43: return false
44: end if
45: end procedure

3.2 Results for LED

By applying Algorithm 1, we determine the weights and heights of all two-round characteristics for LED, PHOTON-80 and PHOTON-128 with \( m \) (as explained in Definition 1) equal to 4, 5 and 6 respectively. The results for LED with both key addition operation and constant addition operation are given in Table 1. Here we present the results for the constant value equal to zero, but the results for other constants are also close to the given numbers. In the table, the numbers on the left (in black) are the results when key addition is used, whereas the numbers in the right (in blue) are the results when constant addition is used.

As explained before (see equation (1)), the characteristics have non-zero \( \text{DP}[k](Q) \) for a fraction \( 2^{m_b-\text{weight}(Q)-\text{height}(Q)} \) of the keys, while for all the other keys is zero. This is equivalent to saying that there is no solution for some of these characteristics for the chosen constant. Then, the expected number of characteristics that have a solution equals to the number of characteristics multiplied by the fraction \( (2^{m_b-\text{weight}(Q)-\text{height}(Q)}) \). For instance, consider the case for 5 active S-boxes, with weight 10 and height 7. Then,
the expected fraction becomes $2^{16-10-7} = 2^{-1}$. From Table 1, we can confirm this as follows: For the key addition operation, there are $2^{5.46}$ characteristics. The expected number of characteristics is computed as $2^{5.46} \cdot 2^{-1} = 2^{4.46}$ when we use the constant value $0x0000$ (written in hexadecimal).

Table 1. Number of characteristics (binary logarithm) for LED S-box (nas: number of active S-boxes).

<table>
<thead>
<tr>
<th>nas</th>
<th>char.</th>
<th>height</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>weight</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>19.55/19.55</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>21.08/21.08</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>21.42/21.42</td>
</tr>
<tr>
<td>15</td>
<td>19.98/19.98</td>
<td>19.90/18.89</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>22.64/22.64</td>
</tr>
<tr>
<td>15</td>
<td>23.65/23.65</td>
<td>24.52/23.52</td>
</tr>
<tr>
<td>16</td>
<td>25.80/24.80</td>
<td>24.98/23.98</td>
</tr>
<tr>
<td>17</td>
<td>26.49/24.49</td>
<td>24.16/21.16</td>
</tr>
<tr>
<td>18</td>
<td>25.75/22.75</td>
<td>21.13/17.12</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>-</td>
</tr>
<tr>
<td>16</td>
<td>25.67/24.67</td>
<td>25.50/23.50</td>
</tr>
<tr>
<td>17</td>
<td>28.45/26.45</td>
<td>27.23/24.23</td>
</tr>
<tr>
<td>18</td>
<td>30.27/27.27</td>
<td>28.16/24.16</td>
</tr>
<tr>
<td>19</td>
<td>31.26/27.26</td>
<td>28.22/23.21</td>
</tr>
<tr>
<td>20</td>
<td>31.35/26.35</td>
<td>27.06/21.06</td>
</tr>
<tr>
<td>21</td>
<td>30.19/24.19</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>19.33/18.33</td>
</tr>
<tr>
<td>17</td>
<td>24.43/22.43</td>
<td>23.98/20.98</td>
</tr>
<tr>
<td>18</td>
<td>28.09/25.09</td>
<td>27.00/23.00</td>
</tr>
<tr>
<td>19</td>
<td>30.83/26.83</td>
<td>29.20/24.20</td>
</tr>
<tr>
<td>20</td>
<td>32.85/27.85</td>
<td>30.70/24.70</td>
</tr>
<tr>
<td>21</td>
<td>34.21/28.20</td>
<td>31.49/24.49</td>
</tr>
<tr>
<td>22</td>
<td>34.87/27.87</td>
<td>31.45/23.45</td>
</tr>
<tr>
<td>23</td>
<td>34.71/26.71</td>
<td>30.20/21.19</td>
</tr>
<tr>
<td>24</td>
<td>33.36/24.36</td>
<td>-</td>
</tr>
</tbody>
</table>

A similar reasoning can be used to compute the expected fraction of characteristics which have a solution for the given constant. To be more precise the fractions are computed as $1/2^n - 1/2^n, 2/2^n, ...$ for the inverse diagonals of the table. For all values in Table 1, this fraction is approximately satisfied, except for the following case: For seven active S-boxes, with weight 16 and height 5. If the key addition operation is used, the number of such characteristics with $DP[k](Q) = 2^{-11}$ will be $2^{7.05}$. From the results of our experiments, there is no such characteristic for constant zero, however, if we consider the constant value $0x0123$, the corresponding result is 1.00, which means that the number of such characteristics is 2. All the results for the constant value $0x0123$ are very close to the one in Table 1 (in blue) for constant zero.

### 3.3 Results for photon-80/20/16 and photon-128/16/16

The results for PHOTON-80/20/16 and PHOTON-128/16/16 when there is a key addition are given in Figure 2 and Figure 3, respectively. In both the figures, the weight of the characteristics are grouped according to the active S-box number. As can be seen from
Fig. 2. Number of characteristics (binary logarithm) of the PHOTON-80/20/16 super box.

Fig. 3. Number of characteristics of the PHOTON-128/16/16 super box.

figures, the highest height for each group is obtained when the weight is minimum. Because of the high computational complexity for Algorithm 1, we did not search the number of keys which have solution when the key addition operation is replaced with constant addition operation. However, it is possible to compute the expected number of characteristics from these results by computing the fractions as described in Section 3.2.

4 Effect of S-box and MDS-matrix for the characteristic distribution

The results of our experiments are given in the previous section (see Section 3). In this section, we compare these results with the existing work which was done to AES [9]

and deduce some useful insight information about the choice of two core components (the S-box and the MDS-matrix) in AES-like round functions.

4.1 Effect of the S-box

The S-boxes are the smallest non-linear components of symmetric-key algorithms and their choice significantly affects the security of the algorithm. As a result, it is a well-studied subject and how to construct a secure S-box is known today. For example, the algebraic degree of the S-box must be high or for any given input difference the distribution of the possible output differences should be close to uniform. This means that, for a four-bit S-box (as in the PRESENT S-box), it is not desirable to have the weight less than two for a non-zero input difference. Similarly, for an eight-bit S-box (like the AES S-box), the weight will be not less than six for a non-zero input difference. When we compare the two super boxes with an equal block size \( n_b \) but with different S-box sizes, we observe higher heights as the S-box size gets smaller. Mathematically this can be explained as follows: Choose two different S-boxes with size \( n \) and \( 2n \) bits (shown as \( S_n, S_{2n} \) respectively), then the best differential probabilities through the S-boxes are given as:

\[
Pr_{n}^{max} = \max_{a,b} \{ Pr(S_n(x) + S_n(x + a) = b) \} = 2^{s-n}
\]

\[
Pr_{2n}^{max} = \max_{a,b} \{ Pr(S_{2n}(x) + S_{2n}(x + a) = b) \} = 2^{t-2n}
\]

for some known \( s, t \) where \( 1 \leq s \leq n \) and \( 1 \leq t \leq 2n \). Then, the differential probabilities for the super boxes \( (S_b) \) are:

\[
Pr_{S_b}^{max} \leq (Pr_{n}^{max})^{B} = 2^{s-n} \left( \frac{n_b}{n} + 1 \right)
\]

\[
Pr_{S_{2b}}^{max} \leq (Pr_{2n}^{max})^{B'} = 2^{t-2n} \left( \frac{n_b}{2n} + 1 \right)
\]

where \( B \) and \( B' \) are the branch numbers of the super boxes constructed using S-boxes of \( n \)-bits and \( 2n \)-bits, respectively. We are interested in the case \( Pr_{S_{2b}}^{max} \leq Pr_{S_b}^{max} \) and aim to find under which conditions it is satisfied. By using the equations (2) and (3), we obtain:

\[
1 \leq \frac{Pr_{S_b}^{max}}{Pr_{S_{2b}}^{max}} \leq \frac{2^{s-n} \left( \frac{n_b}{n} + 1 \right)}{2^{t-2n} \left( \frac{n_b}{2n} + 1 \right)}
\]

taking the logarithm,

\[
0 \leq (s - n)\left( \frac{n_b}{n} + 1 \right) - (t - 2n)\left( \frac{n_b}{2n} + 1 \right)
\]

\[
\frac{(t - 2n)}{2(s - n)} \leq \frac{n_b + n}{n_b + 2n} < 1
\]

This means,

\[
Pr_{S_{2b}}^{max} \leq Pr_{S_b}^{max} \text{ iff } t < 2s.
\]

In practice, this condition is always satisfied for the S-boxes that are designed to resist differential attacks.
As an example, we compare the 32-bit super box of PHOTON-224/32/32 with AES. For PHOTON-224/32/32, the best differential probability for the four-bit S-box is $2^{-2}$ and the branch number is nine. Then for the PHOTON super box the maximum probability is computed as $(2^{-2})^9 = 2^{-18}$. On the other hand, for AES, the best probability for the eight-bit S-box is $2^{-6}$ and the branch number is five, then for the AES super box the maximum probability is $(2^{-6})^5 = 2^{-30}$.

From [9], we know that for a fraction $2^{n_b-(\text{weight}(Q)+\text{height}(Q))}$ of the keys the differential probability is non-zero. Then it is possible to estimate the height for a super box as follows:

$$\begin{align*}
0 &< 2^{n_b-(\text{weight}(Q)+\text{height}(Q))} \leq 1 \\
n_b - \text{weight}(Q) - \text{height}(Q) &\leq 0 \\
n_b - \text{weight}(Q) &\leq \text{height}(Q)
\end{align*}$$

Hence, in order to maximize the height, one should minimize the weight (i.e., use the best differential probability for the super box). Therefore, for the AES super box with five active S-boxes the height should be at least $32 - 30 = 2$. This is verified in the previous work and the maximum height is found as five. By using the same idea, for the PHOTON-224 super box the minimum height is calculated as $32 - 18 = 14$ when nine S-boxes are active.

### 4.2 Effect of the MDS-matrix

In the design of LED and PHOTON, a new method ($d$ times application of a special matrix) is used to generate MDS matrices to obtain more efficient implementations in hardware. We decided to investigate the distribution of characteristics when we change the MDS-matrix to a circulant MDS-matrix. We chose LED as an example since it has a smaller block size, which reduces the computational complexity. As it can be seen from Table 2, (the left column in black represents the original results and the right column in blue shows the results for LED with a circulant MDS-matrix) the distribution of the characteristics does not change. Only the number of characteristics for each cases differs but this is also not that significant. Therefore, we can conclude that the generation method of an MDS-matrix does not change the results of the super box structure.

### 5 Conclusion and Future Work

In this paper, we studied the super box structure of the two recent algorithms, LED and PHOTON. We presented the distribution of all two-round characteristics and also the algorithms that we used. In addition to that, we compared the obtained results and tried to define a new design criterion for the AES-like block ciphers. For a fixed size super box (which is the core of a two-round AES-like structure) the height increases when the S-box size gets smaller. Therefore, it is suggested to increase the number of rounds for the cipher compared with the algorithms with larger S-box sizes.

As a future work, it is possible to extend these results to four rounds covering the \textsc{Step} function of LED. It may not be possible to complete the whole characteristic
Table 2. Plateau characteristics for LED with original / circulant MDS-matrix (nas: number of active S-boxes)

<table>
<thead>
<tr>
<th>nas</th>
<th>char. weight</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>12.97/13.15</td>
<td>4.46/6.85</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>16.92/17.02</td>
<td>9.81/11.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>19.55/19.51</td>
<td>13.85/14.72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>21.08/20.98</td>
<td>16.71/16.81</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>21.42/21.44</td>
<td>18.56/18.28</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>19.98/20.41</td>
<td>18.89/18.95</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>-</td>
<td>-</td>
<td>16.14/15.88</td>
<td>12.03/11.66</td>
<td>3.91/5.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>20.14/20.02</td>
<td>17.07/16.94</td>
<td>10.82/10.78</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>22.64/22.63</td>
<td>20.84/20.83</td>
<td>15.88/15.83</td>
<td>7.55/7.53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>23.65/23.71</td>
<td>23.52/23.56</td>
<td>20.00/19.89</td>
<td>13.37/13.06</td>
<td>2.58/</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>24.80/24.87</td>
<td>22.98/22.92</td>
<td>17.92/17.57</td>
<td>8.60/6.58</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>24.49/24.50</td>
<td>21.16/20.99</td>
<td>14.21/13.13</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>22.75/22.63</td>
<td>17.12/17.38</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>18.21/18.24</td>
<td>16.86/16.86</td>
<td>12.50/12.64</td>
<td>4.95/6.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>24.67/24.66</td>
<td>23.50/23.52</td>
<td>19.33/19.47</td>
<td>12.29/12.52</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>26.45/26.43</td>
<td>24.23/24.25</td>
<td>18.89/18.98</td>
<td>9.96/10.75</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>27.27/27.26</td>
<td>24.16/24.16</td>
<td>17.54/17.57</td>
<td>6.67/7.58</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>27.26/27.26</td>
<td>23.21/23.22</td>
<td>14.92/15.02</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>26.35/26.36</td>
<td>21.06/21.10</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>22.15/22.19</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>18.33/18.41</td>
<td>17.73/17.74</td>
<td>14.03/14.08</td>
<td>7.24/8.23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>22.43/22.41</td>
<td>20.98/20.97</td>
<td>16.15/16.27</td>
<td>8.65/9.76</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>25.09/25.07</td>
<td>23.00/23.02</td>
<td>17.00/17.15</td>
<td>8.61/8.00</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>26.83/26.83</td>
<td>24.20/24.22</td>
<td>16.90/17.08</td>
<td>5.83/6.60</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>27.85/27.88</td>
<td>24.70/24.72</td>
<td>15.87/16.10</td>
<td>5.32/3.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>28.20/28.20</td>
<td>24.49/24.50</td>
<td>13.48/13.92</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>28.18/28.18</td>
<td>23.45/23.46</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>26.71/26.71</td>
<td>21.19/21.17</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>24.36/24.36</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

distribution for the STEP function (due to the high computational complexity) but we believe that this method can be used to obtain characteristics with high height which can be used in an analysis.

References

Abstract. In the last decade, customized services have become widely-used in online applications as they offer personalized products and services to people. These services mostly rely on generating recommendations based on the data collected from the customers, which is often considered as sensitive. To protect the privacy of this data without degrading the service quality number of cryptographic-based solutions were developed. Unfortunately, existing solutions targets only the case of single recommendation computation and do not consider the threats, which appears in the dynamic scenarios with multiple service invocation. In this paper we aim to study the new hazards brought by the dynamic scenarios and develop tools for defending against them.

Keywords: Privacy, recommender systems, cryptography, homomorphic encryption, multiparty computation.

1 Introduction

In the last decade, customized services have become widely-used in online applications as they offer personalized products and services to people. These services mostly rely on generating recommendations using collaborative filtering techniques [4], where the idea is to find similar users and generate recommendations based on their likes and dislikes, which heavily relies on the data collected from the users implicitly or explicitly [5]. The gathered information is personal and thus, privacy sensitive, which creates serious privacy risks due to a number of reasons: theft due to inadequate security measures, re-purposing data for commercial use, and transferring data to third parties [6].

To protect users private data against curious entities, including the service provider and recommendation receiver, without degrading the service quality, cryptographic techniques have been applied [4,5]. Unfortunately, the existing works only address generating recommendations in a static environment, where the number of users does not change in time. Even though these solutions provide provable privacy protection in static settings, their sequential invocation with changing number of users leaks information, damaging the purpose of the privacy-preserving protocol.

In this paper, we aim to address dynamic environments that are de-facto standard for the real-world online services, such as online social networks (OSN) [7], where a number of users joins and leaves the system at each moment. We propose a novel method for preserving privacy of user data in such environments and introduce the cryptographic protocols implementing this method in two different security models, namely semi-honest model with and without collusion. The method is based on the idea of randomly excluding few users from the recommendations computation. We present a few different approaches for excluding users to address the settings with different security and performance requirements.

We develop the protocols, which implements the proposed method and its modifications. The protocols are designed to work in a multiparty setting where a number of users participate in the computations. The protocols are based on cryptographic tools, namely additively homomorphic encryption [8], and applies techniques known from multiparty computation [9]. Our complexity analysis shows that the proposed protocols are efficient in terms of computation and communication. This encourages to deploy these protocols for a realistic recommender system with privacy protection.

The groups with dynamic participation have drawn attention in the cryptographic community, especially to solve the problem of key management [10,11]. There is also prior work in data publishing to protect the privacy of users in case of continuous publishing of data of dynamic user groups [12,13]. Nevertheless, to the best of our knowledge, there has been no previous work addressing dynamic settings for recommender systems.
The rest of this paper is structured as follows. In Section 2, we formalize the new class of attacks feasible in dynamic settings and introduce the method for protecting against them. We also describe the modifications of this method, targeting either higher security or higher performance. In Section 3, we present the cryptographic protocols implementing the protection method and its modifications in two different security models, while Section 5 contains the cryptographic primitives used for these protocols. In Section 6, we provide an overview of proposed protocols complexity and conclude this paper in Section 6.

2 Proposed solution

In this section we introduce the new class of attacks on privacy-preserving recommender systems feasible in dynamic settings and propose the method for protecting privacy of user data under that attacks.

In our settings, a user represents a party that holds a constant private data — preference vector, used during evaluating recommendations. All users are numbered and denoted as $U_i$, their private data is denoted as $d_i$. Users can participate in the group, named the user group $G$. This group can be requested to suggest a recommendation for the given query (which is also a preference vector), denoted as $q$. For simplicity we assume that there is the only one party, denoted as $A$, which can submit queries to $G$. Users from $G$ process $A$’s queries in a privacy-preserving fashion using one of the aforementioned cryptographic protocols and return resulting recommendation $r$. These protocols guarantees that $A$ will learn only the value of $r$ and nothing more about users’ private data.

Next we consider more practical scenario, when $A$ can submit his query few times, say twice. It is clear, that between this two submissions the user group may change — few users can leave or join it. To be more precise, let us consider the following scenario:

1. $A$ submits the query $q$ to the group $G$, in which users $U_{i=1,...,N}$ are involved, and receive the computed recommendation $r_1$.
2. The new user $U_{N+1}$ joins the group $G$.
3. $A$ resubmits the same query $q$ to $G$ and receive the new recommendation $r_2$, which was computed using data from old $N$ users and the new user $U_{N+1}$.

From the pair $(r_1, r_2)$ $A$ can learn a lot of information about private value of $U_{N+1}$. For example, if $r_1 = r_2$ $A$ conclude that preference vector $d_{N+1}$ is not similar to the queried preference vector $q$ or (less likely) that $d_{N+1} = r_1 = r_2$.

We denote the scenarios like above as the new group attacks. In the rest of the paper, we develop security mechanisms against this kind of attacks based on group masking method:

**Method (Group masking).** On each recommender system execution, in which only users $U_{i=1,...,N}$ are involved, the vector $e \in \{0, 1\}^N$, denoted as a group mask, is randomly selected and recommendation is computed using only the data of users $U_i$ for which $e_i = 1$:

$$r = R(q, \{d_i \mid 1 \leq i \leq N, e_i = 1\}),$$

(1)

where $R(\cdot, \cdot)$ denotes the recommendation function. Generated $e$ should be kept secret from $A$.

It is clear that the group masking method hides information about what users are involved in the recommender system execution from $A$. Hence, $A$ cannot link computed values $r_j$ with used $d_i$, and hence it cannot conduct the new group attack.

We propose three different approaches for generating a group mask, each having its own advantage:

1. $e$ is generated in such way that it involves a pre-defined number of users. That is, there exists a publicly known function $Q(\cdot)$, such that $\sum_{i=1}^{N} e_i = Q(N)$.
2. $e$ is generated in such way that each $U_i$ has a pre-defined probability to be involved. That is, there exists a publicly known $p$, such that $P(e_i = 1) = p$.
3. $e$ is generated in such way that it involves at least a pre-defined number of users. That is, there exists a publicly known function $Q'(\cdot)$, such that $\sum_{i=1}^{N} e_i \geq Q'(N)$.

As Approaches 1 and 2 generate group masks with the higher a priori knowledge about the result, this approaches provide more information to potential attackers than Approach 3, and thus they are less secure. Nevertheless, Approaches 1 and 2 have their own advantages, which make them preferable in certain
scenarios: Approach 2 can be implemented much more efficiently than Approaches 1 and 3, and thus it introduces a tradeoff between complexity and privacy. Approach 1 has one advantage over other two approaches — it generates group masks with a pre-defined number of involved users, which is significant for generating recommendations with guaranteed level of quality.

In the next sections, we show how the group masking method, with masks generated according to Approaches 1–3, can be implemented using cryptography. For the sake of simplicity, we do not apply the group masking method for the problem of computing recommendations, but apply it for the simplified problem — for aggregating users’ private data. User data aggregation can be formalized as follows: each user $U_i \in \{1, \ldots, N\}$ holds the private value $d_i \in \mathbb{Z}$; party $A$ should receive $r$:

$$r = \sum_{i=1}^{N} e_i d_i,$$

where $e_i$ are the elements of the group mask vector $e$, generated using one of the approaches above. Party $A$ should not learn any extra information about $e_i$ and $d_i$ than it can deduce from $r$; users should not learn neither $r$ nor other users’ $d_i$.

### 3 Preliminaries

In this paper, we assume the semi-honest security model, which states that all parties participating in the protocol follow the protocol steps correctly, but are curious. Curious parties can collect all data it observes to extract more information than it is allowed to have.

We consider two variations of the semi-honest model:

A. Semi-honest without collusion: parties cannot collude with each other, i.e. each party has access only to its own private data and received messages.

B. Semi-honest with collusion: parties can form coalitions. All colluding parties share their observations in attempt to disclose the private values of other parties. We do not consider the case when the coalition involving all parties may exist, i.e. we assume that the number of users participating in each coalition is upper-bounded by $N - 1$.

In both models, we assume that the parties communicate over an authenticated channel, and though the protocol is protected against an external adversary.

To make our protocols secure and privacy-preserving in aforementioned privacy settings, we use the cryptographic tool named the Paillier encryption [15, 16] which is asymmetric, additively homomorphic and semantically secure. Additive homomorphism allows to compute from encryptions of two messages, say $x$ and $y$, the ciphertext which decryption yields the sum of these messages:

$$[x] \cdot [y] = [x + y],$$

where $[\cdot]$ denotes the encryption functions. Consequently, one can compute $[xc] = [x]^c$.

As the Paillier cryptosystem is semantically secure, one plaintext can be encrypted to different ciphertexts depending on a random factor. Any party, owing only the public key of the encryption scheme, can build a ciphertext equivalent to the given one. This operation is named re-randomization and is denoted as Rand ($\cdot$).

We additionally use the threshold version of Paillier cryptosystem [15] (its modification [17], which works without a trusted dealer, can be used instead). In $K$-out-of-$N$ threshold cryptosystem, there is no single party holding the private key, and thus being able to perform a decryption, but there are $N$ parties holding shares of that key, and contribution from any $K$ of them is required for decrypting. We denote the threshold Paillier encryption function as $J^{\cdot K}$.

### 4 Protocols

In this section, we describe a number of protocols that apply Approaches 1–3 in the semi-honest security settings with or without collusions. The described protocols involve only users $U_{i=1, \ldots, N}$ and the party $A$, which are connected with broadcasted and point-to-point secured channels. Protocols are designed under the assumption that the value $\sum_{i=1}^{N} d_i$ is small enough to be encrypted with Paillier encryption.
Input: Each $U_i$ holds his private value $d_i$. 
Output: Party $A$ receives $r = \sum_{i=1}^{N} d_i e_i$.

1. Each $U_{i=2,...,N}$ encrypts his private data and sends resulting $[d_i]$ to $U_1$.
2. $U_1$ selects uniformly random group mask $e \in_R \{v \in \{0,1\}^N \mid \sum_{i=1}^{N} v_i = Q(N)\}$.
3. $U_1$ encrypts his private data, computes:
   \[ [r] = \left[ \sum_{i=1}^{N} d_i e_i \right] = \prod_{i=1}^{N} [d_i]^{e_i} = \prod_{i=1}^{N} [d_i] \]
   and transfer the result to $A$.
4. $A$ decrypts the received value and gets the resulting $r$.

**Protocol $P_1$: Data aggregation with 1-st group masking, non-colluding setting.**

### 4.1 Protocols for semi-honest setting without collusion

Following protocols preserve the users’ privacy by hiding their data from all other users and party $A$, and $A$ privacy by hiding computation result from all users. We protect the protocols from the new user attack by using group masking, where group mask should be kept secret from $A$.

We assume that Paillier encryption scheme has already been set up: private decryption key $sk$ is passed to $A$, and corresponding public encryption key $pk$ is known to all parties. In the following protocols all encryptions are done using $pk$.

The protocols implementing Approaches 1–3 for the security settings with no collusion are similar with each other. Hence, due to the space limitation, we present only one protocol — protocol with group masking using Approach 1 and explain how two others can be derived from it. The protocol is described in Protocol $P_1$.

Next we sketch the proof of privacy and security of the protocol. On Step 1 $U_1$ receives the encryptions of other users’ private data, which leaks nothing about the values of $d_i$ due to the semantic security of used Paillier encryption and the fact, that $U_1$ cannot access the decryption key. On Step 2 $U_1$ learns the value of $e$. This does not damage the security of the protocol, since $U_1$ cannot infer any information about $r$ or $d_i$ from value $e$. Computations on Step 3 obviously cannot leak anything to $U_1$, due to used encryption. And finally, on Step 4 $A$ receives $[r]$, which is its own private output.

The only modification needed to adopt Protocol $P_1$ or using Approach 2 or 3 is to change the way how the group mask is generated by $U_1$ on Step 2 to what is specified by the selected approach. Note that modified versions are vulnerable to timings attacks [18], as the number of operations executed on Step 3 depends on the $\sum_{i=1}^{N} e_i$. Hence, this value can be indirectly revealed by observing the total computation time of $U_1$.

### 4.2 Protocols for semi-honest setting with collusion

In this section we describe the protocols, which are secure in the semi-honest settings, where parties can form a coalitions involving at most $N - 1$ users. Hence, in this settings we have only one additional security requirement to thus stated in Section 4.1 generated group mask should be kept hidden from all parties, but not only from $A$.

We assume that the Paillier $N$-out-of-$N$ threshold encryption scheme has already been set up: its private key has been shared between all users (each user receives share $sk_i$) and its public key $pk$ are known to all parties. In the following protocols all encryptions are done using $pk$.

**Protocol with group masking using Approach 1** First we present the protocol, which employs the proposed group masking method, where the mask is generated according to Approach 1. The generated group mask $e$ should have the following property: $e \in_R \{0,1\}^N$ and exactly $Q(N)$ its components are equal to 1. Next $Q(N)$ is denoted as $m$ for notation simplicity.

To generate $e$ following the condition above we use the multiple-try method, which process as follows: (i) users generate $t$ vectors $\beta_j \in_R \{0,1\}^N$ in parallel such, that $\forall l \in [1, N]: P(\beta, l=1) = m/N$; (ii) users
Input: Each $U_i$ holds his private value $d_i$.
Output: Party $A$ receives $r = \sum_{i=1}^{N} d_i e_i$.

1. Users jointly generate $tKN$ random encrypted bits: $j \in [1,t]$, $i \in [1,N]$, $t \in [1,k]$ \([\alpha_{j,i,t}] : \alpha_{j,i,t} \in \{0,1\}\).
2. Users jointly run the bitwise less-than protocol $tN$ times in parallel, computing for each $j \in [1,k]$, $i \in [1,N]$ value $[\beta_{j,i}]$:

   \[ [\beta_{j,i}] = \left[ \alpha_{j,i,1} \ldots \alpha_{j,i,t} < \left\lceil 2^k m/N \right\rceil \right] . \]

3. Users locally compute $[b_j] = \left[ \sum_{i=1}^{N} \beta_{j,i} \right] = \prod_{i=1}^{N} [\beta_{j,i}]$, and jointly run $t$ decryptions in parallel to open values $b_j$ to each other.
4. Each $U_i$ selects minimum $j$, such that $b_j = m$, and set $[e] = [\beta_{j,i}]$. This step fails with probability $2^{-\kappa}$.
5. Each $U_i$ computes $[d_i e_i] = \text{Rand} \left( [e_i]^{\beta_{j,i}} \right)$ and broadcasts the result.
6. Users locally compute $[r] = \left[ \sum_{i=1}^{N} d_i e_i \right] = \prod_{i=1}^{N} [d_i e_i]$, and jointly run decryption to open $r$ to $A$.

Protocol $P_2$: Data aggregation with 1-st group masking, colluding setting.

Select as $c$ the first $\beta_j$, such that its elements sum $b_j = \sum_{i=1}^{N} \beta_{j,i}$ is equal to $m$. Multiple-try method is applicable, because the rate of suitable candidates $\beta_j$, i.e. vectors satisfying $\sum_{i=1}^{N} \beta_{j,i} = m$, is fixed:

\[
S = P \left( \sum_{j=1}^{N} \beta_{j,1} = m \right) = \binom{N}{m} \frac{m^m (N-m)^{N-m}}{N^N} \approx \sqrt{\frac{N}{2\pi m(N-m)}} . \tag{4}
\]

Consequently, by executing the sufficient numbers of tries $t$, we can guarantee that the method will fail only with negligible probability $2^{-\kappa}$, where $\kappa$ denotes the statistical security parameter (usually is chosen around 80 bits). In practice, we can note that $S \geq 1/\sqrt{2\pi m}$, and though we can use the following estimation of the value $t$: $t = \kappa \sqrt{2\pi m \ln 2} \approx [1.74 \kappa \sqrt{m}]$.

To perform described multiple-try approach in a privacy-preserving manner, we should generate each vector $\beta_j$ jointly random, i.e. in a such way that $N$ users contribute to it and any subgroup of users together cannot infer, which of candidates for $\beta_j$ are more likely.

Jointly-random generation of a vector $\beta \in R \{0,1\}^N$ is performed as follows:

1. In the precomputation stage the value $k$ is selected, such that the probability $m/N$ can be estimated with $[2^k m/N]/2^k$, i.e. such that the round-up error is negligible:

   \[
   [2^k m/N]/(2^k m/N) - 1 < 2^{-\kappa} . \tag{5}
   \]

2. For each $\beta_i$ (element of $\beta$) users jointly generate random $k$-bit number $\alpha \in R [0,2^k - 1]$, which is available only in encrypted form. This can be done by executing the jointly random bit generation protocol \cite{19} $k$ times in parallel.
3. Then $\beta_i = \alpha_i < [2^k m/N]$ is computed. Comparison is done under the encryption by employing the bitwise less-than protocol \cite{19} (its optimized version from \cite{20} can be used, as one of the compared numbers is publicly known).

The final protocol is presented in Protocol $P_2$. Note that rerandomization of computed $[d_i e_i]$ on Step 5 is required, because otherwise the fact that $d_i = 0$ can be detected by checking whether $[d_i e_i] = 1$ or not. As for the rest, it is clear, that the overall protocol is privacy-preserving and secure, due to the corresponding properties of used cryptography primitives and subprotocols. Opening of values $b_j$ on Step 3 does not leak any information about $e$, as $m = \sum_{i=1}^{N} e_i$ is a priori knowledge.

Protocol with group masking using Approach 2 Protocol $P_2$ can be simplified to achieve relaxed requirements of Approach 2, where only the probability $p$ of user participation is fixed, but not the total number of participating users. Resulting protocol is presented in Protocol $P_3$. 


Input: Each $U_i$ holds his private value $d_i$.
Output: Party $\mathcal{A}$ receives $r = \sum_{i=1}^{N} d_i e_i$.

1. Users jointly generate $kN$ random encrypted bits: $i \in [1,N], l \in [1,k] \rightarrow \alpha_{i,l} \in \{0,1\}$.
2. Users jointly run the bitwise less-than protocol $N$ times in parallel, computing $[e_i] = [\alpha_{i,k} \ldots \alpha_{i,1} < [2^k p]]$.
3. Each $U_i$ computes $[d_i e_i] = \text{Rand}([e_i]d_i^r)$ and broadcasts the result.
4. Users locally compute $[r] = \left[ \sum_{i=1}^{N} d_i e_i \right] = \prod_{i=1}^{N} [d_i e_i]$, and jointly run decryption to open $r$ to $\mathcal{A}$.

Protocol $\mathcal{P}_3$: Data aggregation with 2nd group masking, colluding setting.

Protocol with group masking using Approach 5 Approach 5 requires the group mask $\epsilon$ to have the following property: $e \in_R \{0,1\}^N$ and $\sum_{i=1}^{N} e_i$ is uniformly random in $[Q'(N),N]$. To generate such $\epsilon$, three steps are executed: (i) uniformly random $r \in_R [0,N-Q'(N)]$ is jointly generated (next $N-Q'(N)$ is denoted as $\sigma$ for notation simplicity); (ii) $r$ is converted to $v \in \{0,1\}^r$, which contains exactly $r$ ones; (iii) $v$ is supplemented by $Q'(N)$ ones and permuted to produce $\epsilon$.

To implement the described protocol steps we need to present two additional protocols: secure unary conversion and jointly random shuffling.

Secure unary conversion. This protocol transforms encrypted integer $[r]$, which is from the interval $[0,\sigma]$, to encrypted vector $[[v]] : v \in \{0,1\}^\sigma$ of the following form:

$$v = (1,\ldots,1,0,\ldots,0).$$

Note, that each vector element $v_i$ can be computed using the function $v_i(x) = (x \leq i)$ as $v_i = v_i(r)$. Admitted region of $v_i(x)$ is $\mathbb{Z}_{\sigma+1}$, and hence, $v_i(x)$ can be evaluated in all possible $\sigma$ points and then represented as a Lagrange polynomial:

$$v_i(x) = \sum_{j=0}^{\sigma} v_i(j) \prod_{l=0, l \neq j}^{\sigma} \frac{x-l}{j-l} = \sum_{j=0}^{\sigma} \mu_i,j x^j. \quad (7)$$

Using the observation above, we can describe the secure unary conversion protocol. First, all users jointly run the prefix multiplication protocol [21] to compute $([r^2],\ldots,[r^\sigma])$ from $[r]$. And then, for each $i \in [1,\sigma]$ the value of $[[v_i]]$ can be computed using the Equation (7).

Jointly random shuffling. This protocol generates a jointly random permutation $\pi$, applies it to a given vector and rerandomizes the result. This protocol is based on matrix representations of permutations.

Matrix form of permutation $\pi = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$ is the full-range matrix $M(\pi) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$. Applying a permutation to a vector and composing two permutations in the matrix form are performed using left-multiplication: $\pi(v)^T = M(\pi) \cdot v^T$ and $M(\pi_2 \circ \pi_1) = M(\pi_1) \cdot M(\pi_2)$. Moreover, if a permutation matrix and a source vector are both encrypted, then applying the secure matrix multiplication protocol [22] to them produces rerandomized shuffle, i.e. $[[M(\pi) \cdot v^T]] = \text{Rand}(\pi([[v]]))^T$.

Consequently, to execute jointly random shuffling, it is enough to jointly random generate permutation matrix. This can be done by multiplying permutation matrices generated by each users independently. To execute such matrix multiplication one can employ the unbound fan-in matrix multiplication protocol [21][22].

1. This protocol can be used, while the matrix size/field size ratio is negligible. In practice, the field size in $2^{1024}$ and value is called negligible, when it is less or equal to $2^{-80}$. Hence, this approach is applicable until $N \leq 2^{344}$, which is quite mild restriction.
Input: Each $U_i$ holds his private value $d_i$.
Output: Party $A$ receives $r = \sum_{i=1}^{N} d_ie_i$.

1. Users jointly run the jointly random bounded number generation protocol [19] and produce $[r] : r \in \mathbb{Z}_{q+1}$.
2. Users jointly run the prefix multiplication protocol and compute $([r^2], \ldots, [r^\sigma])$ from $[r]$.
3. $U_1$ builds and broadcasts vector $[v]$:
   
   $[v] = \begin{cases} 
   [\mu_{i,0}] \cdot \prod_{j=1}^{\sigma} [r^j]^{|i-j|} & \text{if } i \in [1, \sigma], \\
   [1] & \text{if } i \in [\sigma+1, N],
   \end{cases}$

4. Each $U_i$ generates random $N$-dimensional permutation $\pi_i$, and builds $M(\pi_i)$.
5. Users jointly run the unbound fan-in matrix multiplication protocol, computing combined permutation matrix $[M(\pi)] = \prod_{i=1}^{N} M(\pi_i)$.
6. Users jointly run the secure matrix multiplication protocol and compute $[e] = [M(\pi) \cdot v^T]^T$.
7. Each $U_i$ computes $[d_i e_i] = \text{Rand}([e_i]^{d_i})$ and broadcasts the result.
8. Users locally compute $[r] = \left[ \sum_{i=1}^{N} d_i e_i \right] = \prod_{i=1}^{N} [d_i e_i]$, and jointly run decryption to open $r$ to $A$.

**Protocol $P_4$: Data aggregation with 3rd group masking, colluding setting.**

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Number of rounds</th>
<th>Number of exponentiations</th>
<th>Number of bits transferred</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>2</td>
<td>$O(N)$</td>
<td>$O(N)$</td>
</tr>
<tr>
<td>$P_2$</td>
<td>13</td>
<td>$O(kN^2 \sqrt{Q(N)})$</td>
<td>$O(kN^2 \sqrt{Q(N)})$</td>
</tr>
<tr>
<td>$P_3$</td>
<td>12</td>
<td>$O(kN^2)$</td>
<td>$O(kN^2)$</td>
</tr>
<tr>
<td>$P_4$</td>
<td>22</td>
<td>$O(N^5)$</td>
<td>$O(N^4)$</td>
</tr>
</tbody>
</table>

By combining two described protocols we can obtain the protocol, which implements Approach 3 for group mask generation. We present the protocol in Protocol $P_4$. The protocol is secure and privacy-preserving, because all underlying subprotocols and cryptographic primitives are secure and privacy-preserving. The protocols’ Steps [13] and Steps [45] should be executed in parallel to reduce an overall round complexity and execution time.

5 Complexity analysis

To give an idea of the efficiency of designed solutions, we give an asymptotic estimation of three measures having the most significant influence on the performance: number of executed interactive rounds, number of bits passed through the network and number of exponentiations executed locally (an exponentiation in the encrypted domain is the most complex local operation; effect of all other operations is negligible comparing to it). The complexities is presented in Table 1.

6 Conclusion

In this paper we propose a method to provide protection of user data processed by a group service in dynamic scenarios, which are more realistic than static ones for a wide range of applications. This method is realized using a set of cryptographic protocols, which are designed with performance in mind. The protocols are shown to be correct, secure and privacy-preserving. The complexity analysis with respect to the versions in two attacker models clearly shows the advantages and disadvantages of the protocols in terms of computational and communication costs, and the level of privacy protection. Our protocols can be further used as building blocks for implementing privacy-preserving recommender systems in a dynamic setting.
References

Precoder Optimization for DSL with High Crosstalk

Arnaud Martin, Luc Vandendorpe, Jerôme Louveaux
Université Catholique de Louvain, Louvain-la-Neuve, Belgium
jerome.louveaux@uclouvain.be

1 Introduction

Next generation DSL systems aim to provide very high data-rates by transmitting at much higher frequencies. Current transmissions go up to 20MHz while transmissions up to 100MHz are considered for G.Fast. At such high frequency ranges, the electromagnetic coupling between twisted pairs of a binder group, named Crosstalk, becomes very large, sometimes even larger than the direct signal transmitted on the line. Efficient crosstalk precompensation techniques (also called precoding) are thus necessary in order to provide the targeted bit rates. The method exploiting the colocation of the transceivers or receivers in order to cancel the crosstalk is called Vectoring [3]. Various precoding techniques have already been designed for current transmission standards (using a frequency range up to 20-30 Mhz) but all the methods considered up to now use the assumption that Far-End Crosstalk (FEXT) is significantly lower than the signal received by the user. The goal of this paper is to propose and analyze different crosstalk cancellation techniques and compare their efficiency in the special case of very high crosstalk, applicable in the high frequency range.

2 System model and parameters

The signal model represents a downstream transmission from a Central-Office (CO) to $N$ different users. The model we used for this work is made on the assumptions of synchronization of the modems and discrete multi-tone (DMT) modulation. The model can then be independently represented for each tone:

$$y = H x + n$$  \hspace{1cm} (1)

Where $x, y \in \mathbb{C}^{N \times 1}$ are respectively the vectors of sent and received symbols, $H \in \mathbb{C}^{N \times N}$ is the channel matrix and $n$ is the vector of additive white Gaussian noise with variance $\sigma^2$. 


The main figure of merit used to compare the algorithms is the spectral efficiency for each tone which is directly linked to the SNIR at the receiver $i$ and to the SNR-gap to capacity $\Gamma$ through the Shannon relation.

$$R_i = \log_2 \left( 1 + \frac{SNIR_i}{\Gamma} \right)$$

(2)

The data-rate $DR_i$ for user $i$ on this sub-carrier is then simply the spectral efficiency multiplied by the symbol rate.

The parameters used in order to simulate the desired transmissions depends on the protocol in work. The VDSL standard is shown as a reference, our interest here is to characterize a transmission for the G.Fast standard, not yet definitely defined.

<table>
<thead>
<tr>
<th></th>
<th>VDSL</th>
<th>G.Fast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duplexing Division</td>
<td>FDD</td>
<td>TDD</td>
</tr>
<tr>
<td>Number of DMT tones</td>
<td>4096</td>
<td>2048 (or 4096)</td>
</tr>
<tr>
<td>Frequency Range (MHz)</td>
<td>0.2 − 17.7</td>
<td>2 − 108 (or 214)</td>
</tr>
<tr>
<td>Tone Width (kHz)</td>
<td>4.3125</td>
<td>51.75</td>
</tr>
<tr>
<td>Symbol Rate (kHz)</td>
<td>4</td>
<td>48</td>
</tr>
<tr>
<td>$\Gamma$ (dB)</td>
<td>12.9</td>
<td>10.8</td>
</tr>
<tr>
<td>Maximum spectral efficiency (bps/Hz)</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Signal Power per Line (dBm/Hz)</td>
<td>−60</td>
<td>−76</td>
</tr>
<tr>
<td>White Noise Power (dBm/Hz)</td>
<td>−145</td>
<td>−140</td>
</tr>
</tbody>
</table>

The per-tone channel matrix used to simulate the transmission comes from the characterization of a 24 pairs cable. This characterization has been done for 1600 frequencies between 100kHz and 250MHz. The magnitudes of the first line elements of the channel matrix are represented on Figure 1 as an example.

Because this cable is relatively short, 50m, the channel matrices below 100MHz have a high diagonal dominance and are then very beneficial for any precompensation algorithm. We have then decided to consider only subcarriers of the 100 − 200MHz bandwidth to simulate the transmission.

3 Analysis of existing technique

The first objective is to identify the efficiency of existing crosstalk cancellation techniques for higher frequencies.
In this idea, modelling of per tone capacity has been made for the classic linear Zero-Forcing Crosstalk canceller [1]. This technique enables us to entirely cancel the FEXT components for each user by defining a linear precoder \( P = H^{-1}H_{diag} \). The outputs of the precoder are \( x = P\beta_{ZF} u \) where \( u \) is the vector with the end-user data symbols.

\[
y = H P\beta_{ZF} u + n
\]  

A particular attention has to be made on the power enhancement created by the Zero-Forcing constraint. After precoding, the symbols have to be scaled down in order to satisfy the power constraint on the transmitted PSD at each subcarrier (4). This is done by imposing a common scaling factor \( \beta_{ZF} \) (5) to all symbols from the different lines. Simulation results show that when the crosstalk is high, this scaling factor can become low and generate significant noise enhancement, making the simple zero-forcing technique less efficient.

\[
\begin{align*}
\beta_{ZF}^2 (|P_{11}|^2 + \cdots + |P_{1N}|^2) &\leq 1 \\
\vdots \\
\beta_{ZF}^2 (|P_{N1}|^2 + \cdots + |P_{NN}|^2) &\leq 1
\end{align*}
\]  

\[
\beta_{ZF} = \left( \max_i \sqrt{\sum_{j=1}^{n} |P_{ij}|^2} \right)^{-1}
\]
4 Improvement of the Zero-Forcing technique

A way of improving the zero-forcing method is to change the scaling of the entire matrix by a scaling of each symbol individually before the linear precoding which does not affect the ZF property of the precoder. Mathematically, it is represented by the right-multiplication of the precoder by a diagonal matrix $G$.

$$\mathbf{P}_{\text{norm}} = \mathbf{H}^{-1}\mathbf{H}_{\text{diag}} \mathbf{G}$$

$$\mathbf{G} = \begin{bmatrix} g_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & g_N \end{bmatrix}$$

The $i^{th}$ output of the precoder can be defined as $\sum_{j=1}^{N} P_{ij} g_j u_j$ and the PSD constraints can then be characterized.

$$\begin{cases} |P_{11}|^2 g_1^2 + \ldots + |P_{1N}|^2 g_N^2 \leq 1 \\
\vdots \\
|P_{N1}|^2 g_1^2 + \ldots + |P_{NN}|^2 g_N^2 \leq 1 \end{cases}$$

The symbols received at users side are found as (8), where $\mathbf{H}_{\text{diag}} \mathbf{G}$ is diagonal and represents the transmission gains. The weights have to be chosen carefully in order to stay in the power constraints while maximizing the sum of the rates.

$$\mathbf{y} = \mathbf{H}_{\text{diag}} \mathbf{G} \mathbf{u} + \mathbf{n}$$

Equal power per user

A first possibility is to assign the factors $g_i$ depending on the vertical norm of the precoder. A global normalization is then applied on the entire matrix in order to bound the maximum line power. The normalization is then done $Per User$ and the same power is given to all of them.
\[ g'_j = \frac{1}{\sum_{i=1}^{n} |P_{ij}|^2} \]  \hspace{1cm} (9)

\[ \beta_{ZF,G} = \left( \sqrt{\max_i \sum_{j=1}^{n} |(PG)_{ij}|^2} \right)^{-1} \]  \hspace{1cm} (10)

\[ g_j = g'_j \ast \beta_{ZF,G} \]  \hspace{1cm} (11)

### Optimized power distribution

A second and very interesting normalization technique is to determine the optimal $G$ matrix, maximizing the sum-rate $\sum_{i=1}^{N} R_i$ while satisfying the constraints (7).

If we assume high SNIR and that $SNIR_{noG}$ is the SNIR when the $G$ matrix is not applied at the precoder, the function to maximize becomes:

\[ \sum_{i=1}^{N} R_i = \sum_{i=1}^{N} \log_2 \left( 1 + \frac{SNIR_i}{\Gamma} \right) \approx \sum_{i=1}^{N} \log_2 \left( \frac{g_i^2 SNIR_{noG,i}}{\Gamma} \right) \]  \hspace{1cm} (12)

The optimization is assured if we maximize the product $\prod_{i=1}^{N} g_i^2$ while respecting the PSD constraints (7). This optimization problem is convex and can be solved exactly with a numerical method called, known as the interior point method.

### 5 Simulations

Simulations have been made in order to compare different techniques on the chosen frequency range. As explained earlier, the subcarriers of the 100 – 200 MHz frequency range have been considered from the cable characterization.

If we take the sum of the spectral efficiencies for all frequencies and multiply by the symbol rate, we get the total data-rate for each user on the entire transmission bandwidth, represented on Figure 2.
By comparison to the scalar normalization of the linear zero-forcing precoder, the average data-rate improvement per line is 13.8 Mbps for the Equal Power per User normalization, representing approximately 4% of the average data-rate, and is 55 Mbps for the Optimized normalization, representing 15.5% of the average data-rate. The improvement is then relatively important considering that only the normalization has been modified.

For next analyses, we hope to find a normalization algorithm that would be closer to the optimized normalization efficiency but easier in term of numerical complexity.

The second analysis to be made is to determine from which degree of row-wise diagonal dominance the new normalization techniques become interesting to use. This degree is defined for each subcarrier $k$ by the parameter $\alpha_k$, which has a theoretical expression (13) for its maximum value [2].

$$|h_{n,m}^k| \leq \alpha_k|h_{n,n}^k| \quad \forall m \neq n$$

$$\alpha_k \leq K_{fext} f_k \sqrt{d_{coupling}}$$

(13)

Where $h_{ij}^k \in H_k$, the channel matrix for the subcarrier frequency $k$, $f_k$ is the frequency on tone $k$ in MHz, $d_{coupling}$ is the coupling length in kilometers and $K_{fext} = -22.5 dB$.

By representing the data-rate difference between new techniques and classical scalar normalization for the entire bandwidth of the characterization, $0.1 - 250 MHz$, on
Figure 3 and comparing it to the row-wise diagonal dominance of the channel matrix with the frequency on Figure 4, we can determine from which $\alpha_k$ parameter value the new algorithms become preferable.

![Figure 3: Differences of data-rates with the linear ZF and scalar norm. technique](image)

![Figure 4: Theoretical and experimental degree of diagonal dominance of channel matrix](image)

From this comparison, we can see that both normalizations become preferable from 100 MHz, where the $\alpha_k$ parameter becomes higher than 0.15.
6 Conclusion

This document investigates the efficiency of the linear precoder at high frequencies for a transmission channel characterized on a 24 pairs cable. The regular scalar normalization for the linear zero-forcing precoder is near-optimal at low frequencies but becomes less efficient as the frequency becomes higher.

By improving the normalization process, it is possible to improve this technique while keeping a very similar precoding algorithm. The normalization is done per user while respecting PSD constraints and maximizing the sum-rate. The new normalization techniques have shown better performance, especially at high frequencies and specifically when the degree of row-wise diagonal dominance surpass the extracted value.

References


[2] Transmission and Multiplexing (TM); Access transmission systems on metallic access cables; VDSL; Functional Requirements, ETSI Std. TS 101 270-1, Rev. V.1.3.1, 2003.

Equalization of the non-linear 60 GHz channel: Comparison of reservoir computing to traditional approach

M. Bauduin*, T. Deleu*, F. Duport†, P. De Doncker*, S. Massar‡, F. Horlin*

*Université Libre de Bruxelles - OPERA Wireless Communication department
mbauduin@ulb.ac.be
†Université Libre de Bruxelles - OPERA Photonique department
‡Université Libre de Bruxelles - LIQ department

Abstract

The non-linearities in a communication channel can severely affect the communication quality. These problems are encountered in many communication systems. Because of the high complexity of their power amplifiers, which have a severe non-linear behaviour, combined with an important pathloss, which imposes an important output power, the 60 GHz communications are strongly affected by these non-linearities. Taking these non-linearities into account in channel equalization can increase the communication performances and enable us to work near the saturation point of the amplifier. This paper presents the reservoir computer as a new approach for the equalization of a non-linear communication channels in the case of the 60 GHz communications. We compare the performances and the complexity of the reservoir computer algorithm with an iterative maximum likelihood (ML) equalizer. We find that the reservoir computer is an interesting low complexity solution for this task.

I. INTRODUCTION

The capacity requirement in wireless communications is constantly evolving and we reach the performance limits of the present wireless systems. Communications at 60 GHz are an interesting solution because there exists, at this frequency, a large unlicensed available bandwidth [1] which enables communications with important bit rate. Because of the important pathloss, these communications are limited to short distances which enables an important frequency reuse. This high bit rate technology limited to one room can have a lot of applications like high definition uncompressed video streaming, wireless gaming, wireless gigabit Ethernet, data transfer or equipment synchronisation.

This new communication system has a lot of interesting technology challenges, including the channel equalization. Indeed, because of the high bandwidth, the memory of the channel is important and we cannot consider a narrowband channel for the equalization [2]. The conception of 60 GHz amplifiers is also complex due to the high frequency, making the 60 GHz channel non-linear. These non-linearities introduce an additional distortion of the signal which is very penalizing for high order modulations, like the 16-QAM modulation which is the modulation considered in this paper.

This work proposes a new approach for the non-linear channel equalization which consists in using a machine learning algorithm. Specifically we use a bio-inspired algorithm:
the reservoir computing which is derived from the neural networks [3] [4] [5] [6] which can be trained to minimize the mean square error on the estimated symbol sequence. Because of the memory of the reservoir computer and its non-linear characteristic [7], this structure is a good candidate for non-linear channel equalization. Preliminary results on reservoir computing applied to channel equalization were reported in [8] but with a non-realistic channel model. To evaluate the performances of the reservoir computer, we compare it with an iterative algorithm based on the ML criterion.

The outline of this paper is the following. In section 2, we present the channel communication model. In section 3, we consider a linear MMSE (Minimum Mean Square Error) equalizer which can only attenuate the linear interferences. In section 4, we propose an iterative ML equalizer which will be used as a benchmark. The reservoir computer will be presented in section 5. The comparison in performance and complexity of these algorithms will be done in section 6.

II. 60 GHz COMMUNICATION CHANNEL

The baseband communication channel is described in figure 1(a). We consider half-root Nyquist shaping filters. The propagation channel is obtained with a stochastic model proposed in [2]. Here, we will only consider a Line Of Sight (LOS) communication.
channel. This channel represents an important challenge for the equalization because of the important sampling frequency, imposed by the standards, which results in an important memory. This specificity implies the use of equalizers which must also have an important memory.

The power amplifier, which is the main source of non-linearities in most of the communication systems, is given by a non-linear baseband model which characterises the AM-AM and AM-PM characteristic of the power amplifier [9]. If the input is defined as $x(n) = a(n)e^{j\phi(n)}$, we have the following signal, $y(n)$, at the output of the amplifier:

$$y(n) = f(a(n))e^{j(\phi(n)+g(a(n)))}$$  \hspace{1cm} (1)

where $f(.)$ represents the AM-AM relation and $g(.)$ represents the AM-PM relation.

The amplifier we will use is the GaAs pHEMT 60GHz HPA by NEC. We can characterise this amplifier with a Rapp model [10]:

$$f(x(n)) = \frac{G|x(n)|}{1 + \left(\frac{2|x(n)|}{V_{sat}}\right)^{2p}}$$  \hspace{1cm} (2)

$$g(x(n)) = \frac{A|x(n)|^q}{1 + \left(\frac{|x(n)|}{B}\right)^q}$$  \hspace{1cm} (3)

with the following parameters: $G = 19$, $V_{sat} = 1.4$, $p = 0.81$, $A = -48000$, $B = 0.123$, $q = 3.7$.

The AM-AM and AM-PM characteristics of the amplifier are represented in figure 2. The operating point is defined by the OBO (Output Back Off) which is defined by the following equation:

$$OBO = 20\log_{10}\frac{V_{out}}{V_{sat}}$$  \hspace{1cm} (4)

where $V_{out}$ is the mean amplitude of the output signal and $V_{sat}$ is the saturation amplitude of the amplifier.

III. IMPACT OF THE NON LINEARITIES

To evaluate the impact of the non-linearities, we will first equalize the channel with a linear minimum mean square error (MMSE) equalizer (figure 1(b)). We make the hypothesis that the channel is perfectly known by the receiver. But in order to characterise this equalizer, a linear approximation of the channel around an operating point must be done. So the non-linear characteristic of the channel are not considered by the equalizer. The impact of these negligences can be observed in figure 3.

A linear approximation of the channel has its limits and we need equalizers which consider more parameters of the channel. The two equalizers proposed in the next sections will take into account these non linearities, thereby improving the performance of the equalization.
Fig. 2. AM-AM (characterized with the OBO) and AM-PM (characterized with the phase) relation of the power amplifier in function of the amplitude of the input signal

Fig. 3. Linear MMSE equalization of a non linear 16-QAM channel communication with different level of OBO

IV. ML ITERATIVE EQUALIZER

This algorithm is based on the modified order P compensation algorithm [11] adapted for the equalization. The first step of the equalizer consists in a linear equalization to make a first estimation of the transmitted signal ($\tilde{s}_0(n)$ in figure 1(c)). We will use the linear MMSE equalizer for this step. After this first estimation, the algorithm will use an iterative process to estimate the transmitted signal $\tilde{s}_i(n)$ which produced the received one, $r(n)$. This implies that the receiver has a perfect knowledge of the non-linear channel to be able to evaluate which signal $\tilde{r}_i(n)$ the channel will produce with a input $(s)_i(n)$. The objective is to find a sequence $\tilde{s}_i(n)$ which produces an output $\tilde{r}_i(n)$ close to $r(n)$.

At each step $i$, we will modify each symbol, one by one, in the estimated sequence $\tilde{s}_{i-1}(n)$ by adding a $\Delta$ to each symbol. By this way, we move, at each iteration, each
symbol from its initial estimation to a value \( \tilde{s}_i(n) \) which meets the ML criterion

\[
\tilde{s}_i(n) = \tilde{s}_{i-1}(n) + \Delta
\]  

(5)

where \( \tilde{s}_i(n) \) is the estimated symbol for the step \( i \) and \( \Delta \) is the modification.

The objective is to find the distance \( \Delta \) between the actual estimated symbol and its effective position in the emitted constellation. This value of \( \Delta \) is the one which creates a \( \tilde{r}_i(n+l) \) close to \( r(n+l) \) for each \( l \in [-L_c'; L_c - 1] \) where \( L = L_c + L_c' \) is the channel memory. The impact of \( \Delta \) on the sequence \( \tilde{r}_{i-1}(n) \) is evaluated with:

\[
\tilde{r}_i(n + l) = \tilde{r}_{i-1}(n + l) + F_{NL}(\Delta)
\]  

(6)

where \( F_{NL} \) characterizes the non-linear comportment of the channel.

It is very complicated to derive the optimal value of \( \Delta \) from this expression. So we use a linear approximation of the channel to reduce the complexity of the evaluation of \( \Delta \). If \( \Delta \) is defined to minimize the euclidean distance between the sequences \( \tilde{r}_i(n) \) and \( r(n) \), we obtain the following expression:

\[
\Delta = \frac{\sum_{l=-L_c'}^{L_c-1} a_l^*(r(n+l) - \tilde{r}_{i-1}(n+l))}{\sum_{l=-L_c'}^{L_c-1} (a_l a_l^*)}
\]  

(7)

where \( a_l \) is the linear approximation of the channel.

As the channel is known by the receiver, we can estimate the coefficients \( a_l \) by adding a known \( \varepsilon \) to the actual symbol \( \tilde{s}_{i-1}(m) \)

\[
\tilde{s}_{\text{test}}(m) = \tilde{s}_{i-1}(m) + \varepsilon
\]  

(8)

This will produce the signal \( \tilde{r}_{\text{test}}(n) \).

\[
\tilde{r}_{\text{test}}(m) = \tilde{r}_{i-1}(m) + F_{NL}(\varepsilon)
\]  

(9)

The linear approximation of the channel can be done with the following equation:

\[
a_l = \frac{\tilde{r}_{\text{test}}(m + l) - \tilde{r}_{i-1}(m + l)}{\varepsilon}
\]  

(10)

We may not forget that the optimal \( \Delta \) is obtained by a linear approximation of the channel. A too high value of \( \Delta \) would not have the desired impact. This is why the \( \Delta \) can be multiplied by a weight factor \( \mu \in [0; 1] \). With a little value of \( \mu \), the convergence speed is reduced but the linear approximation is more reliable. An important value of \( \mu \) can accelerate the convergence speed but the linear approximation is less reliable. At each iteration, the algorithm will modify each sample of the emitted sequence to find the one which optimizes the ML criterion.

The iterative bloc uses the following schema:

1) Estimate the received sequence \( \tilde{r}_i(n) \) which corresponds to the estimated sequence \( \tilde{s}_i(n) \)
2) Evaluate the coefficients \( a_l \)
3) Evaluate the optimal \( \Delta \)
4) Add the optimal $\Delta$ and return to point 1)

The complexity of this equalizer is $O(M.K.O_{\text{channel}})$ where $K$ is the number of iterations, $M$ is the number of symbols and $O_{\text{channel}}$ is the complexity of the channel evaluation. This last operation is an important source of complexity. For this task, we used 100 iterations and a coefficient $\mu$ of 0.2.

V. THE RESERVOIR COMPUTER

A. Presentation of the reservoir computer

The reservoir computer is an algorithm derived from the neural networks which are automatic learning methods. The memory of this structure is defined by its neurons. The evolution of this system is defined by the following equations [4].

\begin{align}
    a_i(n) &= A \sum_j \alpha_{ij} b_j(n-1) + Bu_i r(n) \quad (11) \\
    b_i(n) &= f(a_i(n)) \quad (12) \\
    \tilde{s}(n) &= \sum_j w_j b_j(n) \quad (13)
\end{align}

where $r(n)$ is the input signal (observed at the channel output), $\tilde{s}(n)$ is the output signal, $b_i(n)$ is the value of the neuron $i$, $\alpha$ is the interconnection matrix, $u$ is the input mask, $w$ is the output mask, $f(.)$ is the connexion function between neurons, $a_i(n)$ is the activation sent to the neuron $i$ at the time step $n$.

The scalars $A$ (feedback gain) and $B$ (input gain) are arbitrary weights which can modify the evolution of the system. The values of $A$ and $B$ will define if the entire system is more influenced by the new input value or its previous state. With an important value of $B$, the system is immediately excited. An important value of $A$ will create a reservoir which stays activated during a long time without excitation. We see that each neuron receives an activation $a_i$ through a function $f(.)$. In general this function is a sigmoid function [3]. The output mask is obtained with an off-line training. A training sequence is sent in the channel and the coefficients of the output mask are evaluated to minimise the mean square error between the estimation of the training sequence and the effective sequence [4].

B. Equalization with a reservoir computing structure

We will use a reservoir computer with the following interconnection matrix:

\[
\alpha = \begin{pmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1 \\
1 & 0 & 0 & \cdots & 0
\end{pmatrix}
\quad (14)
\]

Each neuron is connected to its neighbours in a ring like in figure 4 [12].
Because of the important memory of the channel propagation (around 50 time steps) combined with the non-linearities, the reservoir computer will need too many neurons to compensate all the interferences. To reduce the complexity of a such equalizer, an alternative structure composed by two reservoir computers is proposed (figure 1(d)). The first one will equalize the linear part of the signal. The second one will work in a non-linear regime to equalize the non-linear part.

The first equalizer will work in a linear regime. So its feedback gain will be important (to improve the memory of the reservoir) but to conserve the echo state property of the reservoir computer, this gain must be lower than 1 [4]. In other case, the reservoir can become unstable because the signal sent by each neuron is amplified. The connexions between neurons will be linear. The second one will use an hyperbolic tangent as interconnexion function. To work in a non linear regime, we will use an important input gain and a very low feedback gain. Each reservoir has a complexity of $O(MNO_f)$ where $O_f$ is the complexity of the connexion function $f(.)$. For this task, 300 neurons were used. The output mask has been evaluated with the help of a training sequence made of 10,000 symbols.
An equalizer algorithm based on a reservoir computer structure has been presented to equalize the 60 GHz channel. Because of the important channel memory (which comes from the high frequency sampling) combined with the added non-linearities from the power amplifier, the equalizer should be divided into two parts. The first reservoir computer is configured to work in a linear mode to compensate the important memory of the channel. The second reservoir computer is configured to work in a non-linear regime to compensate the non-linearities. Its performance is compared with the other equalizers on figure 5. This equalizer offers better performances than a simple linear equalizer like the MMSE because the second structure can take into account the constellation mismatching and the non-linear interferences.

This paper was also the opportunity to present an iterative ML equalizer which has the advantage of being able to work with a long channel memory (which is not the case of trellis based equalizer like Viterbi [13]). This structure meets the ML criterion with an iterative process but it requires a perfect knowledge of the channel which is not the case of the two previous algorithms which are working with an approximation of this channel. The simulations show that the iterative ML equalizer offers better performances than the double reservoir computer but this latter has the advantage of keeping a low complexity because the evaluation of an hyperbolic tangent is less complex than the evaluation of the non-linear channel.

REFERENCES

Nonlinear Interference Suppressor for Varying-Envelope Local Interference in multimode transceivers

H.Habibi, E. J. G. Janssen, Wu Yan, P.G.M. Baltus, J. W. M. Bergmans
Eindhoven University of Technology, Department of Electrical Engineerin
Email: H.Habibi@tue.nl, HabibiHooman@gmail.com, Phone: +31402473287

Abstract

In multimode transceivers, a local transmitter may induce a large interference in a local receiver, often several orders of magnitude stronger than the desired received signal. To suppress this interference by linear filtering, the receiver would need a very large dynamic range, resulting in excessive power consumption. A potentially much more power-efficient approach uses an adaptive memoryless nonlinearity that can strongly suppress the interference when adapted proportional to the envelope of the received interference. This approach has so far been limited to constant-envelope interferences owing to the difficulty of extracting accurate interference envelope information from the received signal. In this paper, we observe that in multimode transceivers the locally available baseband interference enables accurate adaptation for varying-envelope interferences. We identify and analyze nonlinear distortion products which are negligible for constant-envelope interferences. We show that adequate interference suppression can be achieved along with a negligible distortion to the desired signal.

I. INTRODUCTION

The number of communication standards supported by handheld devices has been increasing rapidly in recent years. To implement these standards in a single device, a combination of several transceivers is required, which is called a multimode transceiver. Owing to the small size of a handheld device, the transmitted signal of a Local Transmitter (LTX) is received by a Local Receiver (LRX) for another communication standard with a small attenuation, inducing an interference many orders of magnitude stronger than the desired signal in the LRX. For example, let us consider simultaneous operation of a WLAN Receiver (RX) operating in the frequency range of 2400-2483 MHz and a local WiMAX transmitter (TX) operating in the frequency range of 2496-2690 MHz. Power of the transmitted WiMAX signal can be as high as 23 dBm, while power of the WLAN received signal can be as low as -82 dBm [1]. The coupling loss between transceivers in a multimode transceiver is typically between 10 to 30 dB [2]. Hence the locally induced interference by the WiMAX LTX can be as high as 13 dBm, resulting in a Signal to Interference Ratio (SIR) of -95 dB to -75 dB at the input of the WLAN RX Front-End (FE).

In principle, an interference with no spectral overlap with the desired signal can be completely suppressed by linear filtering. A Bandpass Filter (BPF) is typically used after the LRX antenna to suppress the out-of-band interferences. Typically, the interference suppression by the BPF is from 10 to 40 dB, depending on the frequency separation of the desired signal and interference. For the above WLAN LRX plus WiMAX LTX scenario, this suppression results in a SIR of -85 dB to -35 dB after the BPF. If the receiver FE was exactly linear, further filtering could be done after down-conversion of the received
signal. The FE however, has a limited linear dynamic range. Presence of an interference, beyond this range leads to excessive loss of FE gain and hence leads to loss of sensitivity of the LRX. Increasing the dynamic range to handle this strong interference requires an increase in power consumption which is not acceptable for handheld devices [3]. Hence the interference must be suppressed at an early stage of the receiver. An alternative approach to linear filtering is to suppress the interference by passing the input signals through a memoryless nonlinearity [4]. Its input-output characteristic, as shown in Fig. 1, can be realized by combining a limiter with an adaptable limiting amplitude \( l(t) \) and a linear amplifier with gain of \(-c\). We call this a Nonlinear Interference Suppressor (NIS). The NIS input includes an interference much stronger than the desired signal. The limiter gain for the interference is positive and proportional to \( l(t) \) divided by the input envelope. For a constant-modulus interference, \( l(t) \) can be tuned such that the limiter gain for the interference equals to \( c \). Hence the NIS gain for the interference equals to 0 and the interference is suppressed at the NIS output. On the other hand, owing to the compressive behavior of the limiter, the limiter gain for the desired signal is smaller than \( c \). Hence the NIS gain for the desired signal will be strictly larger than 0. An early implementation of the NIS was used in [5] to suppress a strong constant-envelope interference in spread spectrum receivers.

The limiting amplitude \( l(t) \) that results in complete interference suppression depends on the envelope of the received interference at the NIS input. For a constant-envelope interference, \( l(t) \) must be slowly adapted to track the changes in the power of the received interference [6]. For a varying-envelope interference, \( l(t) \) must be adapted proportional to the envelope of the received interference. In multimode transceivers the transmitted baseband interference is locally available. We propose to generate the adaptation signal from the baseband interference, as shown in Fig. 2. The impact of LTX and LRX components on the envelope of the received interference, from the baseband transmitted interference to
the received interference at the NIS input, can be taken into account digitally. Hence in this paper we assume that the adaptation signal \( l(t) \) can be determined accurately. A novel state of the art implementation of the NIS for varying-envelope interferences can be found in [7]. We show that by using the NIS the local interference can be substantially suppressed. Hence the receiver with the NIS will require a linear dynamic range much smaller than that of a receiver without NIS (which we henceforth call the baseline receiver).

We will see that using the NIS for varying envelope interferences leads to introduction of in-band nonlinear distortion products, which are negligible for constant-envelope interferences. These products, which were not identified in previous work [4] [5] [8], are categorized as:

1- Gain Variation Distortion (GVD): The NIS gain for the desired signal depends on the ratio of envelope of the desired signal to envelope of the interference. As a result the gain varies over time and this leads to distortion of the desired signal. The GVD can degrade the Symbol Error Rate (SER) of the receiver. The degradation increases when SIR at the NIS input increases. For these larger SIRs the baseline receiver can handle the interference without the NIS.

2- Inter-modulation (IM) leakage: The IM is centered at a frequency different from the center frequency of the desired signal. Depending on the frequency separation of the desired signal and interference, however, a part of the IM can leak into frequency channel of the desired signal. For the smallest frequency separation of the desired signal and interference this IM leakage can limit the SER performance of the receiver.

II. RECEIVER MODEL WITH NONLINEAR INTERFERENCE SUPPRESSOR

![Fig. 3: Direct conversion receiver with NIS.](image)

Fig. 3 shows a direct conversion receiver with the NIS. The signal collected by the antenna, including both the local interference and the desired signal, is passed through a Band-Pass filter (BPF). The desired signal is passed almost unchanged through the BPF and the interferences is suppressed to some extent by the BPF. Even after the BPF, however, the interference can be many orders of magnitude stronger than the desired signal. The BPF output \( x(t) \) includes both the desired signal and interference as:

\[
x(t) = A_d(t) \cos(2\pi f_d t + \varphi_d(t)) + A_i(t) \cos(2\pi f_i t + \varphi_i(t)).
\]

where \( A_i, \varphi_i, f_i, A_d, \varphi_d, f_d \) are envelope, phase and center frequencies of the interfering and desired signals after the BPF, respectively. The BPF output \( x(t) \) is passed through the NIS to suppress the interference. Average SIR at the NIS input is defined as: \( \text{SIR}_x = \frac{\langle E(A_d^2) \rangle}{\langle E(A_i^2) \rangle} \), where \( \langle \rangle \) denotes statistical expectation. The NIS output \( y(t) \) is amplified by a Low Noise Amplifier (LNA), down-converted by a quadrature mixer, passed through a Low-Pass filter (LPF), sampled and quantized by an Analogue to Digital Converter (ADC).
III. NONLINEAR INTERFERENCE SUPPRESSOR (NIS)

In this section, firstly we derive the adaptation signal that leads to complete interference suppression in the absence of the desired signal. For this adaptation signal we then derive the NIS output in the presence of the desired signal and identify the key distortion products at the NIS output.

A. NIS modeling and adaptation

As shown in Fig. 1, the NIS can be built by combining output \( y_a \) of a linear amplifier and output \( y_i \) of a hard limiter with an adaptable limiting amplitude \( l(t) \). By changing \( l(t) \), we can change the input-output characteristic of the NIS. We are interested in the conditions that the interference is much stronger than the desired signal. Hence we first look at the simple case where only interference is present. In this case the NIS input will be: \( x(t) = A_i(t) \cos(2\pi f_i t + \varphi_i(t)) \). The NIS output \( y(t) = f(x(t)) \) has harmonic components with center frequencies at integer multiples of \( f_i \). We assume that all the harmonic components, except the fundamental component at \( f_i \), will be filtered out in the proceeding stages. Hence, we only consider the fundamental component of \( y(t) \). By using the Fourier series expansion, the fundamental component is obtained as:

\[
y(t) = A_{i,y}(t) \cos(2\pi f_i t + \varphi_i(t)) = \left( \frac{4l(t)}{\pi} - cA_i(t) \right) \cos(2\pi f_i t + \varphi_i(t)). \tag{2}
\]

By solving \( A_{i,y}(A_i) = 0 \), the optimal adaptation signal that nulls the interference at the NIS output is obtained as:

\[
\tilde{l}(t) = \frac{\pi cA_i(t)}{4}. \tag{3}
\]

B. NIS output in the presence of the desired signal

In the presence of the desired signal, the NIS output \( y(t) \) includes three dominant components with center frequencies close to \( f_d \) [9]:

\[
y(t) \cong A_{d,y}(t) \cos(2\pi f_d t + \varphi_d(t)) + A_{i,y}(t) \cos(2\pi (f_d + \Delta f) t + \varphi_i(t))
+ A_{IM}(t) \cos(2\pi (f_d + 2\Delta f) t + 2\varphi_i(t) - \varphi_d(t)), \tag{4}
\]

where \( A_{d,y}(t), A_{i,y}(t) \) and \( A_{IM}(t) \) are envelopes of the interference, desired signal and main Inter-Modulation (IM) component at the NIS output, respectively.

1) Interference suppression: For \( A_i(t) > A_d(t) \), by using a series expansion for the hard limiter output [10] we obtain:

\[
A_{i,y}(t) \simeq \left( \frac{4l(t)}{\pi} - cA_i(t) - \frac{l(t) A_d^2(t)}{A_i^2(t)} \right)_{l(t)=\tilde{l}(t)} = -\frac{c}{4} A_d^2(t) A_i^{-1}(t). \tag{5}
\]

Suppose that Instantaneous SIR at the NIS input and output are defined as: \( \text{ISIR}_x(t) = \left( \frac{A_d(t)}{A_i(t)} \right)^2 \) and \( \text{ISIR}_y(t) = \left( \frac{A_{d,y}(t)}{A_{i,y}(t)} \right)^2 \) respectively. Then using (5) we obtain:

\[
\text{ISIR}_y(t) \simeq 4 \ ISIR_x^{-1}(t). \tag{6}
\]

According to (6) the instantaneous SIR at the NIS output will be about 6 dB larger than inverse of the instantaneous SIR at the NIS input. Hence the local interference, which is stronger than the desired signal at the NIS input, is suppressed such that it would be weaker than the desired signal at the NIS output.
2) **Distortion products:** For $A_i(t) > A_d(t)$, by using a series expansion for the hard limiter output \([10]\) we obtain:

\[
A_{d;g}(t) \simeq \left( \frac{2l(t)}{\pi A_i(t)} - 1 \right) A_d(t) + \frac{l(t)}{4\pi} \frac{A_d^3(t)}{A_i^3(t)} = -\frac{c}{2} A_d(t) + \frac{c}{16A_i^2(t)} A_d^3(t). \quad (7)
\]

\[
A_{\text{IM}}(t) \simeq \left( -\frac{2A_d(t)}{\pi A_i(t)} l(t) \right)_{l(t) = i(t)} = -\frac{c}{2} A_d(t). \quad (8)
\]

Instantaneous gain $g_d(t)$ of the desired signal is defined as $g_d(t) = \frac{A_{d;g}(t)}{A_d(t)}$ and by using (7) it is obtained as:

\[
g_d(t) \simeq -\frac{c}{2} + \frac{cA_d^2(t)}{16A_i^2(t)} = -\frac{c}{2} + \frac{c}{16}\text{ISIR}_x(t). \quad (9)
\]

According to (9), $g_d(t)$ varies over time. The variation of gain leads to in-band distortion of the desired signal. The Gain Variation Distortion (GVD) is a general form of cross-modulation distortion. The cross-modulation is the transfer of interference modulation to the small desired signal and is only a function of $A_i(t)$ \([11]\). According to (9) as ISIR$_x$ decreases the gain approaches a constant value of $-\frac{c}{2}$. Hence the GVD increases as SIR at the NIS input increases. For a constant envelope interference only variations of $A_d(t)$ contributes to the GVD. For a varying envelope interference variations of both $A_d(t)$ and $A_i(t)$ contributes to the GVD. Hence it is expected that using the NIS to suppress a varying envelope interference leads to more GVD compared to that of a constant envelope interference. In Section IV-B, the impact of the GVD in the SER of the receiver is investigated for a variety of modulations.

According to (5) and (8), an IM component with the same envelope as the desired signal will be present at the NIS output. The IM component is a nonlinear mixture of the desired signal and interference with a frequency separation of $2\Delta f$ with respect to the desired signal. Depending on the frequency separation of the desired signal and interference, a part of the IM may leak into frequency channel of the desired signal. In section IV-A, the IM leakage will be evaluated for the WLAN RX plus WiMAX TX scenario.

### IV. Simulation Results

For simulations, we consider the scenario of the WLAN LRX plus WiMAX LTX. The received desired WLAN signal has a center frequency of 2472 MHz and a bandwidth of 20 MHz. The WLAN signal has Orthogonal Frequency Division Multiplexing (OFDM) modulation with 64 sub-carriers, where each subcarrier can have QPSK, 16 QAM or 64 QAM modulation. The transmitted WiMAX signal occupies the frequency range of 2496-2690 MHz with bandwidth of 10 MHz. We consider two center frequencies for the WiMAX signal: 2502 MHz and 2532 MHz, resulting in frequency separations of $\Delta f = 30$ MHz and $\Delta f = 60$ MHz. We consider two cases for WiMAX signal modulation: constant-envelope modulation and OFDM modulation.

#### A. Evaluation of IM leakage

The IM component is the largest component with small frequency separation from the desired signal. A part of the IM component may leak into frequency channel of the desired signal. Fig. 4 shows a numerical evaluation of the power ratio of the desired signal to the IM
leakage vs. $\Delta f$ for the WLAN RX plus WiMAX TX scenario. The WLAN and WiMAX signals both are OFDM modulated and have rectangular shaped frequency spectrums. The power of IM leakage in 20 MHz bandwidth of the WLAN signal is measured by simulation. We observe that the amount of IM leakage decreases 9 dB by doubling $\Delta f$ when $\Delta f$ is large. The IM leakage adds to the channel noise and it can degrades the SER.

![Graph showing power ratio of desired signal to IM leakage vs. $\Delta f$ for WLAN RX and WiMAX TX scenario.](image)

**Fig. 4:** Power ratio of desired signal to IM leakage vs. $\Delta f$ for WLAN RX and WiMAX TX scenario.

**B. SER comparison of the baseline RX and the RX with NIS**

We assume that the received WLAN signal is passed through an Additive White Gaussian Noise (AWGN) channel. Hence the SER performance of the baseline RX depends only on the desired Signal to Noise power Ratio (SNR), where the noise power is measured in the frequency channel of the desired signal. The SNR is chosen such that it results in an un-coded SER of $10^{-3}$ for the baseline RX. The required SNR for QPSK, 16 QAM and 64 QAM is 10.3, 17.6 and 24 dB, respectively [12]. On the other hand, because of the GVD and IM leakage, the SER of the RX with the NIS depends on the SNR, SIR$_x$, and $\Delta f$.

1) **SER performance for constant-envelope interference and OFDM desired signal:** Consider the case that the interference has a Gaussian Minimum Shift Keying (GMSK) modulation and the desired signal has an OFDM modulation. In Fig. 5a and Fig. 5b the SER for the RX with the NIS vs. SIR$_x$ is shown for $\Delta f = 30$ MHz and 60 MHz, respectively. In both figures we see that by decreasing SIR$_x$, SER decreases and reaches $10^{-3}$, i.e. SER of the baseline receiver. The SER degradation owing to the GVD depends on SIR$_x$ and becomes evident in both figures when SIR$_x$ increases. The GVD limits the largest SIR$_x$ for which the NIS offers a negligible SER degradation. We can use Fig. 5 to determine this limit for a certain amount of SER degradation. No IM leakage is observed for the constant-envelope interference.

2) **SER performance for OFDM modulated desired signal and OFDM interference:** Now consider the case that both the desired signal and the interference have OFDM modulations. In Fig. 6a and Fig. 6b the SER for the RX with the NIS vs. SIR$_x$ is shown for $\Delta f = 30$ MHz and 60 MHz, respectively. Both figures show that by decreasing SIR$_x$, SER decreases and reaches a floor.

Similar to constant-envelope interference case, the SER degradation due to the GVD becomes evident in both figures when SIR$_x$ increases. However, the observed GDV for an
OFDM interference is much larger than for a constant-envelope interference. The GVD limits the largest SIR\textsubscript{x} which for the NIS offers a negligible SER degradation. Fig. 6 can be used to determine this limit for a certain amount of SER degradation. For example when ∆f = 60 MHz, if we want to keep the SER less than 2 \times 10^{-3} (equivalent to an SNR degradation less than 0.5 dB) then we should stop using the NIS when SIR\textsubscript{x} is larger than -12 dB, -18 dB and -27 dB for QPSK, 16 QAM and 64 QAM, respectively. Based on this simulation we can find a threshold on SIR\textsubscript{x} to use the NIS within a certain amount of SER degradation. When SIR\textsubscript{x} is larger than the threshold the baseline receiver can handle the interference without the NIS aid.

The distance of the SER floor from the ideal SER of 10^{-3} is very small for ∆f = 60 MHz as we see in Fig. 6b. This SER floor, which is independent of SIR\textsubscript{x}, originates from the IM leakage and decreases by increasing ∆f from 30 MHz (Fig. 6a) to 60 MHz (Fig. 6b). The amount of degradation due to the IM leakage can be calculated using Fig. 4. For example for ∆f = 30 MHz, the IM leakage power is 28 dB smaller than the desired signal power. For 16 QAM the SNR to achieve an SER of 10^{-3} is 17.6 dB. Hence the ratio of the desired signal to noise plus IM leakage will be 17.2 dB. This 0.4 dB degradation to
the SNR translates into an SER floor of about $2 \times 10^{-3}$, when GVD becomes negligible (for small SIR$_x$), as we see in Fig. 6a. For $\Delta f = 60$ the IM leakage power becomes 37 dB smaller than the desired signal power and the amount of SNR degradation decreases to 0.05 dB which results in an SER floor of $1.1 \times 10^{-3}$ as we see in Fig. 6b.

V. CONCLUSION

In multimode transceivers, the transmitter for one communication standard may induce a large interference in the receiver for another one. Owing to the limitations of linear analog filtering, the interference can still be several orders of magnitude larger than the desired signal at the input of the receiver front-end. To process the desired signal in the presence of such strong interference a receiver with a large dynamic range and high power consumption is required. A much more power efficient approach is to use an adaptive Nonlinear Interference Suppressor (NIS) which was only used for constant-envelope interferences in the previous works. To enable application of this circuit for varying-envelope interferences in multimode transceivers, we proposed a new adaptation method which exploits the availability of the transmitted interference. We showed that the adaptation method can strongly suppress the interference such that it will normally be much smaller than the desired signal at the NIS output. We identified the main distortion products introduced by the NIS, namely Gain Variation Distortion (GVD) and Inter-Modulation (IM) leakage. The GVD increases when desired Signal to Interference power Ratio (SIR) at the NIS input increases. For larger SIRs the linear receiver without the NIS can handle the interference without requiring an excessive dynamic range and power consumption. The IM leakage is only considerable for smallest frequency separation of the desired and interfering signals and it vanishes rapidly by increasing the frequency separation. Hence for most conditions of practical interest sufficient interference suppression is achieved with negligible distortion of the desired signal.

REFERENCES

Towards perfectly rejecting one dominating harmonic interference in SDR receiver

Chunshu Li, Min Li, Sofie Pollin, Marian Verhelst, Liesbet Van der Perre
Interuniversity Micro-Electronics Center (IMEC) vzw
Kapeldreef-75, Leuven, B-3001, Belgium
E-mail: {chunshu, limin, pollins, marian, vdperre}@imec.be

ABSTRACT

Odd-order harmonic mixing is an important problem in wide-band transceivers with switching mixers. In contrast to the traditional “analog mixing-path recombination” solution which suffers from the limited harmonic rejection performance due to phase and gain mismatch among the mixing paths, this paper proposes to recombine the mixing paths digitally, where near-ideal mismatch correction can be performed. An iterative on-die estimation of mismatch and an optimal recombination based on the estimated mismatch for both the transmitter and receiver are presented. The simulation results show that assuming four mixing paths, a single dominating harmonic interferer can be fully rejected for transceivers with more than 15dB initial harmonic rejection ratio under realistic radio input scenarios.

Keywords: Harmonic Rejection, Digital Intensive, Software Defined Radio, Weighted Least Squares

INTRODUCTION

Software Defined Radios (SDRs) allow to transmit or receive any band of interest over a wide frequency range, and hence require wide-band transceiver design. The required flexible up- and down-mixing is commonly implemented using switching mixers [1]. One big challenge of this approach is the harmonic mixing problem: Odd-order harmonics caused by the switching mixer will up-mix baseband signals in the transmitter (Tx) to the radio frequency (RF) band at multiples of the target frequency band, hence causing mask violations. Similarly, RF interfering signals present at multiples of the receive frequency band will be down-mixed to the baseband in the receiver (Rx), distorting the desired signal [2].

Traditional designs either use multiple parallel dedicated single-band RF filters or an RF tracking filter, as proposed in [3] and [4] to filter out harmonic interferers (HIs), which is bulky and power hungry. Multi-path mixing is a promising alternative solution to handle odd-order harmonic mixing [5]-[8]. In this approach, the outputs of multiple switching mixers are combined, each weighted with an appropriate weighting factor to approximate the aggregate LO as a pseudo-sinusoid signal (Fig.1). The closer the aggregate LO signal approximates a sinusoid, the fewer harmonics it contains. As shown in Fig.1, LO₂ and LO₃ are typically 45° and 90° shifted duplicates of LO₁ and each of them contains many odd-order harmonics. It has been shown that an aggregate LO with an exact weight of √2 for LO₂ rejects the 3rd and 5th order HIs completely [2].

To reject HIs down to the transceiver noise floor, normally 60 to 100dB harmonic rejection (HR) ratio is required [5]. However, the achievable HR performance of multi-path mixing solution greatly depends on the phase and amplitude accuracy in each path. Phase and gain mismatch in practical implementations typically limits the HR ratio to 30-40dB [5]. Mismatch compensation has been applied to mitigate this problem. Solutions proposed in [6]-[8] scale the gain of each mixing path by scaling the tail current source in each mixing transistor and compensate the mismatch with lots of additional analog components (LO buffers, resistors and capacitances). All these approaches perform mismatch mitigation and signal recombination in the analog domain, which suffers from limited mismatch calibration accuracy, implementation difficulties and area inefficiency.

The design proposed in [5] adopts a mixed analog-digital approach to achieve high HR robust to mismatch, which conducts a preliminary HR in the analog domain and further enhances HR performance with digital adaptive filtering. It presents the best HR performance so far to the authors’ knowledge. However, two drawbacks exist in this solution: 1.) The HR performance highly depends on the signal-to-interferer ratio (SIR) of RF input and the initial mismatch among the mixing paths; 2.) Analog recombination circuits are still required in each mixing path to ensure sufficient rejection of the HI before entering the digital domain, resulting in additional area overhead. For simplicity, the analog recombination and mixed-signal compensation method proposed in [5] will be abbreviated to ARMC in this paper.

Recently, in [2], a multi-path mixing receiver architecture has been proposed, which does not require any analog calibration, yet promises beyond state-of-the-art HI suppression. The introduced digital recombination HR receiver recombines the mixing paths’ outputs optimally in the digital domain, based on a priori knowledge of the phase and amplitude in each mixing path. However, up to now, no good solution existed to estimate mixing path mismatch impairments on-chip for either Tx or Rx. This paper proposes a digital intensive HR framework which efficiently estimates phase and gain mismatch among the mixing paths, for both the Tx and Rx, and then computes the optimal digital recombination weights for the mixing paths based on the
mismatch estimation results. To the authors’ knowledge, this is the first work to explore the simultaneous mismatch calibration for HR in both the Tx and Rx. All the estimation and compensation is performed digitally, which benefits from more accurate weighting and relaxed analog matching requirements. The simulation results for four paths show that: A single dominating HI can be rejected completely if no other radio impairments (e.g. LO phase noise) are present, regardless of the input SIR, as long as the initial phase and gain mismatch ensure an HR ratio above 15dB under realistic radio scenarios, hence significantly surpassing performance achievable by analog recombination schemes.

As mentioned before, the proposed calibration methodology aims to jointly estimate the required pre- and post-compensation weights towards optimal mixing path combination in presence of analog phase and gain impairments. The calibration process firstly conducts on-chip mismatch estimation, which is performed at any user-defined time instance, e.g. during system setup or after operational frequency shifts, and then optimally recombine mixing paths during normal transmission based on the estimated mismatch parameters. The remainder of this paper is organized as follows: Section 2 presents the detailed methodology for phase and gain mismatch estimation. Section 3 explains the optimal recombination scheme for a single dominating HI during normal transmission; Section 4 demonstrates the achieved performance; and Section 5 concludes the paper.

**Iterative Phase and Gain Mismatch Estimation for Both Tx and Rx**

This section presents the detailed calibration architecture and schedule for estimating the phase and gain mismatch jointly in Tx and Rx.

Fig. 2 illustrates the proposed architecture for the mixing-path mismatch estimation. It is assumed that either the Tx, the Rx, or both, adopt a multi-path mixing scheme for HR. Equidistant 45° shifted LOs (0° – 45° – 90° – 135°) are provided to the four paths of the Tx and Rx, taking into account an unavoidable phase error $\Delta \theta_{(T/R)n}$ for every LO. The inherent gain of each
path is $G_{(T/R)n}$. In case of a multi-path mixing Tx, weighting for each path is conducted digitally through pre-distortion. For the multi-path mixing Rx, the down-converted baseband signals are, after low pass filtering, directly converted into the digital domain by the ADCs, after which they are appropriately weighted and digitally recombined. In both cases, multi-path recombination in digital domain improves flexibility and accuracy. Note that both phase and gain mismatches are corrected by simple digital weighting.

During the mismatch estimation, the Tx and Rx work in loop-back mode, meaning the Tx output is directly coupled with the Rx input via the parasitic coupling in the transmit-receive switch of the antenna interface, as shown in Fig.2. This mismatch estimation is performed at pre-determined time instances.

The iterative estimation process, shown in Fig.3, works as follows:

**Step 1-3:** All the four paths in Tx are activated with inputs that are digitally weighted with an initial weighting ratio $1 : \sqrt{2} : 1 : 0$ applied to both the in-phase (I) and quadrature-phase (Q) components. In absence of phase and gain mismatch, as shown in the upper graph of Fig.4, the aggregate I and Q signals at the Tx output ($TX_I$ and $TX_Q$) can be perfectly reconstructed without any HI residual. Taking into account the practical phase-shift errors, as shown in the bottom graph of Fig.4, the fundamental signal components in $TX_I$ and $TX_Q$ are distorted by each other and will contain residual harmonic component. The larger the phase or gain mismatch, the more pronounced the presence of the residual. This distorted Tx signal will now be used to calibrate the receiver.

In the Rx under calibration, only two paths are activated in each step, first Rx1 and Rx3 (in step 1), followed by Rx2 and Rx4 (step 2), and then Rx1 and Rx2 (in step 3), as shown in Fig.3. This is possible thanks to the flexibility of digital recombination scheme. In each step, the original multi-path mismatch calibration is reduced to the calibration of traditional IQ imbalance, for which many practical solutions exist [9][10]. Note that joint Tx and Rx IQ-imbalance estimation is required here, to separate the Tx impairments (not under calibration in these steps 1-3) and the Rx impairments under calibration.

**Step 4:** Mismatch estimation among four paths can subsequently be derived from the mismatch estimates between each pair of paths of step 1-3. Based on the recombination scheme explained in Section 3, an updated weighting ratio is determined, which reduces the harmonic leakage in the Rx.

**Step 5-7:** Subsequently the Tx impairments will be calibrated by activating all four Rx paths with their updated digital weighting factors. Similar to step 1-3 for the Rx mismatch estimation, now, pairs of paths in the Tx (Tx1 and Tx3 in step 5, Tx2 and Tx4 in step 6, Tx1 and Tx2 in step 7) are activated two by two. Again, multi-path mismatch calibration is reduced to joint Rx-Tx IQ-imbalance calibration, using [9][10].

**Step 8:** The weighting ratio for the Tx multi-path recombination can be updated based on the Tx mismatch estimation derived in step 5-7. The updated weighting ratio will reduce the Tx harmonic leakage, which will improve the accuracy of the
Step 1-8 are repeated until the predefined performance requirement is met or little additional improvement can be achieved. Note that residual of harmonic distortion in the (T/R)x output \(((T/R)x)_I\) and \((T/R)x)_Q\) will degrade the phase and gain mismatch estimation for each activated pair of paths in (R/T)x. The proposed iterative scheme therefore updates the weighting factors according to the estimated mismatch values in each iteration and reduces the residual of harmonic distortion step by step to gradually improve mismatch estimation accuracy.

EXPLORATION OF OPTIMAL RECOMBINATION

This section describes the mixing-path recombination scheme, which works in two time instances: 1.) During normal transmission, the weighting factor for each path in Tx and Rx should be generated based on the estimated mismatch parameters determined in the previous section; 2.) As described in previous section, both step 4 and step 8 of our iterative calibration scheme, require to compute an update of the digital weighting factors based on new mismatch estimation coefficients.

As shown in Fig.5, depending on which harmonic should be suppressed (e.g. 3rd order), different weighting factors should be applied. This section describes the mixing path recombination strategy with the case of rejecting the 3rd order HI.

Building on the developed mathematical framework presented in [2], the coefficients for path recombination for interference estimation can be derived by Eqn.(1)

\[
\begin{bmatrix}
\text{Cancelling harmonics in I path} \\
\text{Cancelling image frequencies in I path} \\
\text{Cancelling harmonics in Q path} \\
\text{Cancelling image frequencies in Q path}
\end{bmatrix}
\]

which leads to:

\[
\begin{bmatrix}
\sum_{n=1}^{N} S_{In} F_{LOnm} \\
\Re(\sum_{n=1}^{N} S_{In} F_{LOnm}) \\
\sum_{n=1}^{N} S_{Qn} F_{LOnm} \\
\Im(\sum_{n=1}^{N} S_{Qn} F_{LOnm})
\end{bmatrix} =
\begin{bmatrix}
0 \\
0 \\
0 \\
0
\end{bmatrix} \quad \forall m \text{ concerned}
\]

(2)

where \(\Re()\) denotes the real part of a complex number, \(\Im()\) denotes the imaginary part of a complex number, \(F_{LOnm}\) denotes the complex Fourier coefficient of LO’s \(m_{th}\) harmonic in the \(n_{th}\) mixing path, \(S_{In}\) and \(S_{Qn}\) are digital weighting factors for the \(n_{th}\) mixing path for I and Q, \(N\) is the total number of mixing paths and \(m\) represents the order of the HI that need to be rejected (3rd in this case).

Based on the absolute phase \(\theta_{1-4}\) and gain mismatch \(G_{1-4}\) estimations among the four paths, \(F_{LOnm}\) can be written as Eqn.(3).

\[
F_{LOnm} = P_{m,n} \cdot e^{-jm\theta_n} \cdot (1 - (-1)^m)
\]

(3)

where \(P_{1,n} = 2/\pi \times G_n\), \(P_{3,n} = -2/(3\pi) \times G_n\) and \(P_{5,n} = 2/(5\pi) \times G_n\).
Combining Eqn.(2) and Eqn.(3) for the case being explored: (four paths for rejecting the 3\textsuperscript{rd} order HI), gives a solvable set of equations (Eqn.(4)):

\[
\begin{bmatrix}
0 \\
0 \\
0 \\
0
\end{bmatrix}
= 
\begin{bmatrix}
\sum_{n=1}^{4} S_i \times P_{3,n} \cos \theta_n \\
\sum_{n=1}^{4} S_i \times P_{3,n} \sin \theta_n \\
\sum_{n=1}^{4} S_q \times P_{1,n} \cos \theta_n \\
\sum_{n=1}^{4} S_q \times P_{1,n} \sin \theta_n
\end{bmatrix}
\]

A set of scaling factors ($S_{i(1-4)}$ and $S_{q(1-4)}$) can always be found to exactly meet Eqn.(4). This conclusion applies for rejecting any single HI. This means that, ideally, the single (3\textsuperscript{rd} order) HI can be rejected completely if the phase shifts and gain mismatch can be estimated accurately. This will be illustrated in Section 4.

RESULTS AND COMPARISON

Fig.6 shows the simulation performance of our proposed digital intensive HR scheme for the 3\textsuperscript{rd} order HI’s rejection. We will compare its performance with the state-of-the-art ARMC method in [5], which adopts mixed analog-digital rejection solution and presents the best HR performance so far to the authors’ knowledge.

Fig.7 depicts the sketch of the state-of-the-art ARMC HR system framework, where digitally adaptive interference cancelling technique is applied to further improve the HR of the analog mixing-path recombination. In the digital domain, the primary input, containing the desired signal and multiple harmonic residual distortions, is constructed by a linear recombination of the mixing paths. To eliminate the harmonic distortions in this primary signal, adaptive interference cancelling is performed. To this end, a reference input containing the interference estimation is generated by a second linear recombination of the mixing paths. The working mechanism of the adaptive filtering method is to adaptively adjust the amplitude and phase of the interference estimation to produce an output that is as close a replica as possible to the distortion components in the primary input. This output is then subtracted from the primary input to produce the desired signal. Assuming the SIR at the receiver input, the interference estimation input and the receiver output are denoted as $SIR_{in}$, $SIR_{est}$ and $SIR_{out}$ respectively. It is proven in [?] that the maximum achievable $SIR_{out}$ is determined by $SIR_{est}$ as follows:

\[
SIR_{out} = \frac{1}{SIR_{est}}
\]

Assuming the HR ratio of analog recombination is denoted as $HR_{ana}$ and the rejection ratio of the desired signal in the interference estimation is denoted as $DR_{dig}$, the following relationship exists:
Fig. 8. Simulation performance for a single (3rd-order) HI’s rejection using iterative calibration with 12-bit quantization

\[ SIR_{in} + HR_{ana} - DR_{dig} = SIR_{est} \]  \hfill (6)

Combining Eqn.(5) and Eqn.(6), the achievable system HR ratio \((SIR_{out} - SIR_{in})\) of ARMC scheme can be represented by Eqn.(7):

\[ HR = SIR_{out} - SIR_{in} = -2 \times SIR_{in} + DR_{dig} - HR_{ana} \]  \hfill (7)

As a result, the HR performance of this approach highly depends on the SIR of the RF input, as reported in [5], in which 98dB HR could only be theoretically achieved with \(SIR_{IN} = -46dB\). The method introduced in this paper can under equal assumptions achieve >60dB HR in both the Tx and Rx after the second iteration of mismatch estimation. After 15 iterations, the harmonic rejection performance is even perfect (the upper bound in Fig.6 occurs due to the accuracy of the Matlab simulation tool). Besides, the HR performance of the proposed method has no dependency on the RF input.

Moreover, it is robust against large initial mismatch reducing the original HR ratio. A tolerance down to 15dB initial HR is achieved, as shown in Fig.6, which can greatly relax the requirements on the analog design. On the other hand, in [5], the HR in analog domain and the interference generation in digital domain are both blind to the intial mismatch, because no mismatch estimation effort is conducted. Hence, the initial mismatch will influence the values of HR\(_{ANA}\) and DR\(_{DIG}\), and further affect the overall HR ratio that can be achieved.

To show the proposed HR scheme works with realistic quantization errors in Tx and Rx, the quantization accuracy in Tx and Rx is changed to 12-bit enob. Fig.8 shows the simulated HR performance under the realistic quantization. After three iterations, over 85dB HR in both the Tx and Rx can be stably achieved, which is more than enough to reject HIs down to the transceiver noise floor.

Besides the performance improvement, compared to the analog recombination scheme of [5]-[8], the proposed method saves analog circuit design effort and area and is robust to technology scaling by moving multi-path recombination to the digital domain. To the authors’ knowledge, this is the first work to explore the joint mismatch calibration for HR in both the Tx and Rx.

**CONCLUSIONS**

This paper proposes a digital intensive method for HR in both the transmitter and receiver, which saves analog circuits design effort and is robust to technology scaling with digital multi-path recombination. The introduced calibration scheme can theoretically reject a single dominating HI completely regardless of the RF input or initial mismatch. The resulting digital intensive HR transceiver with built-in calibration enables low-cost, high-HR SDR transceivers.

**REFERENCES**


Digital Self-Interference Cancellation for Full Duplex Communication on USRP

Tom Vermeulen, Sofie Pollin
KU Leuven Departement Elektrotechniek - ESAT
Kasteelpark Arenberg 10 postbus 2440, Leuven, B-3001, Belgium
E-mail: tom.vermeulen1@student.kuleuven.be, sofie.pollin@esat.kuleuven.be

Abstract

Full duplex communication starts to receive a lot of attention since it promises to improve the spectral efficiency and coexistence performance of wireless LAN communication. To enable it in practical systems, digital interference cancellation techniques are needed that allow to subtract the self-interference from the wanted received signal. To enable this, accurate channel estimation of the self-interference channel in all conditions is needed. While the state-of-the art focuses on estimating the self-interference channel without interference, or in the time domain only, this paper presents an accurate frequency domain channel estimation for full duplex OFDM communication. The proposed technique achieves a digital self-interference cancellation that is 10dB more accurate compared to standard OFDM LS frequency domain channel estimation, and 2dB better compared to the state-of-the art full duplex interference cancellation in the time domain.

1 Introduction

Since the introduction of wireless LAN more than a decade ago, its use has been accelerating constantly. The Wireless Broadband Alliance estimates that the number of private WiFi hotspots in 2013 is around 492 million[1]. Because all these devices use the crowded ISM bands, solutions to improve coexistence and increase spectral efficiency are needed. The concept of full duplex, which means transmitting and receiving on one frequency with a single device, can help achieve this. In theory the throughput can be doubled and the frequency use can be halved, hence spectral efficiency is improved. Experiments in [2] show that full duplex on average gives an 84% improvement in throughput compared to a traditional half duplex node.

Equally importantly, the use of full duplex facilitates better coexistence in wireless LAN networks, since it promises to solve the hidden node problem. This hidden node problem arises when a node is unable to hear an ongoing transmission and starts to transmit, causing a collision at the receiver of the first transmission. One way to solve it is using Request-to-Send and Clear-to-Send packets, resulting in a high overhead because of the extra packets, and hence barely used. When using full duplex nodes, any receiver can become a transmitter as well, and by doing so inform its environment that it is actively engaged in communication. As a result, any ongoing transmission is equally well detected in the environment of the receiver and the transmitter, and the hidden node problem is eliminated. Now the hidden terminal problem can only arise in
the small period during which the receiving node is decoding the header and starting its transmission [2].

Since the invention of wireless communication it was believed that it is not possible to send and receive on the same frequency at the same time. This mainly comes from the fact that the self-transmitted signal is much stronger than the signal from other nodes[2], saturating the wireless receiver and making it unable to receive any other lower power signal. Figure 1 shows the full duplex block diagram. The self-interference(SI) signal is the one from antenna T1 to antenna R1 and from antenna T2 to antenna R2. Since the T1 is much closer to R1 than T2, the receiver signal power from T1 is 60 to 90dB higher compared to the received signal power from T2. However, if the receiver enables sufficient dynamic range, it becomes possible to subtract this so called SI signal from the received waveform, and leaving enough power in the residual waveform so that the received signal can be decoded[3].

![Figure 1: Full duplex block diagram][8]

There are three ways to cancel the SI: analog cancellation, antenna cancellation and digital cancellation. While the first two will be briefly introduced, the main contribution of this paper is in improving the digital cancellation for the case of asynchronous full duplex communication. In asynchronous operation, it is a challenge to estimate the SI channel \( h_{SI} \) with sufficient accuracy in all scenarios, even when the preamble is corrupted by interference from the other ongoing packets. To generate our results, we implemented the algorithms on a software radio platform taking into account a real front-end, synchronization and baseband implementation.

Frequency domain(FD) channel estimation is the default channel estimation used for OFDM. It is hence also used to calculate the communication channel between Node 1 and Node 2. It could also be used for estimating the SI channel. However, there are two key differences between the communication and SI channel: (a) the SI channel has a very strong line-of-sight and hence little multipath effects resulting in a very fiat frequency response, and (b) the SI channel suffers from noise and interference during the preamble affecting the estimate. As a result, a straightforward application of the same channel estimation algorithm for both channels might not give the optimal results. The novel technique proposed in this paper uses the full packet instead of only the preamble to estimate the channel. This method averages out more of the noise from the other node.

This paper is organized as follows. In section 2 the state of the art will be discussed with a focus on digital cancellation. Section 3 will present two asynchronous full duplex scenarios, and the solutions in this paper will focus on the most challenging one. In section 4 the proposed SI channel estimation will be discussed. Section 5 will present the experimental results of the techniques discussed in this paper, including the software radio implementation approach. The last section, section 6, concludes this paper and suggests some future work.
2 State of the art

To make full duplex possible, it is necessary to cancel the SI signal. The difference with conventional interference cancellation techniques is that the SI signal is known and can be used for the cancellation. Currently there are three techniques used for SI cancellation. Figure 2 gives an overview of these techniques, and shows that they are complementary and can be combined. Each of them is discussed in more details below.

![Figure 2: Block diagram of a wireless full-duplex node. From: [2]](image)

**Analog cancellation** can be done with an analog noise cancellation circuit like the Quellan QHX220[4]. It takes the transmitted signal as reference and subtracts this in the analog domain from the received signal. The Quellan chip, as seen in figure 2, has two inputs. The first input comes from the receive antenna and contains both the useful as the SI signal. The second input comes from the transmit antenna and contains only the SI signal. At its output, the QHX220 delivers a signal with the SI signal cancelled out. Experiments in [5], [2] and [3] show that this chip can reduce a significant part of the SI signal. A reduction of 30dB is achieved in [5] for narrowband signals. [2] achieves a 13dB reduction on a 15MHz signal with the same chip.

**Antenna cancellation** uses two transmit antennas, which transmit the same signal. Figure 2 shows this setup. The distance between the transmit antennas and the receive antenna is different for both antennas. If the wavelength of transmission is equal to \( \lambda \) and the distance between the first transmit antenna (TX1) and the receive antenna is equal to \( d \), the second transmit antenna (TX2) is placed at \( d + \lambda/2 \). This will cause destructive interference at the receive antenna because both transmit signals will add destructively[2]. A 20dB cancellation of the SI signal is possible according to [2]. The experiments in [2] also show that the destructive interference is not only created at the receive antenna but also at other places, this attenuation is around 6dB. One big downside of antenna cancellation is the need for three separate antennas. A 2X2 MIMO system for example uses also two antennas and can in theory also double the throughput when using spatial multiplexing. Although full duplex can solve the hidden terminal problem, its additional gains compared to other multiple antenna solutions are not that pronounced any longer.

**Digital cancellation** works on the digital samples that come from the ADC. If a full duplex radio has a good estimation of the SI channel, it can be used to generate a digital representation of the SI signal from its transmitted signal, and subtract it from the received signal. Experiments in [2] and [3] show a 10dB and 30dB cancellation of the SI signal respectively. The 20dB difference comes from the fact that [3] estimates the SI channel and then generates the digital samples using this channel estimate.
While [2] subtracts the samples without using a channel estimate. Digital interference cancellation is not a new technique; successive interference cancellation (SIC) [6] and ZigZag decoding [7] are examples of techniques that try to solve the hidden terminal problem by iteratively removing interference or collisions. Both techniques use digital cancellation to recover packets that would otherwise be lost. When a SIC implementation can recover 80% of the packets, this is a great success, since that results in an 80% performance gain when recovering packets that were otherwise lost. When a full duplex implementation can only recover 80% of its packets, this system is not useful, and results in a severe performance degradation compared to another wireless link with a typical packet error rate of below 1%. In this respect, digital cancellation for full duplex is much more difficult [3], even if it can benefit from the known transmitted signal.

The techniques discussed in this section can cancel up to 30dB on their own. To cancel the SI signal in an 802.15.4 system for example, more than 60dB of the SI signal will need to be cancelled [2]. In [3] the combined techniques can cancel up to 73dB over a bandwidth of 40MHz, this would enable an 802.11n implementation. These results are achieved in ideal situation where the SI channel is estimated with enough accuracy. This will not always be the case, as explained in Section 3.

3 Asynchronous operation

In wireless lan systems, communication is asynchronous. Whenever a node has a packet to send, it will attempt to send this whenever the channel is found idle. Because full duplex nodes can send and receive at the same time, there are two situations possible. The first situation is the Receive-While-Send (RwS) scenario, where Node 1 is transmitting a packet when it starts receiving a packet from Node 2 after a random time \( \delta \) (Figure 3). If \( \delta \) is larger than the length of the preamble, then the preamble of Node 1 is free of interference from the packet of Node 2, and can be used to estimate the SI channel. The received samples of the preamble are equal to \( y[n] = h_{SI}[n]x[n] + z[n] \), with \( x[n] \) the transmitted signal and \( z[n] \) noise. The channel estimate \( h_{SI} \) can be calculated using standard channel estimation algorithms, because the only noise present is equal to \( z[n] \).

\[
\text{Node 1: } \begin{array}{c|c}
\text{Preamble} & \text{Data} \\
\hline
\text{Node 2: } & \begin{array}{c|c}
\text{Preamble} & \text{Data} \\
\hline\end{array}
\end{array}
\]

Figure 3: Receive-While-Send scenario

This scenario is most widely used in the state of the art, as it brings least challenges to the channel estimation and hence SI cancellation.

However, [8] shows that the Send-While-Receive (SwR) scenario can be up to 1dB worse in BER than the RwS scenario. In the SwR scenario, Node 1 will first start receiving a packet from Node 2, before sending its own packet to Node 2 after time \( \delta \) (Figure 4). The preamble of Node 1, used to estimate the SI channel, will have interference from the packet of Node 2. The received samples of the preamble are in this scenario equal to \( y[n] = h_{SI}[n]x_1[n] + h_{2\rightarrow 1}[t]x_2[n] + z[n] \) with \( x_1[n] \) and \( x_2[n] \) respectively the samples transmitted from Node 1 and Node 2, where \( h_{2\rightarrow 1}[n] \) is the wireless channel between Node 2 and Node 1. The channel estimation will in this scenario be less accurate because of the presence of the \( h_{2\rightarrow 1}(t)x_2[n] \) term, which has to be removed.

The focus will lie on this scenario because of the increased complexity in estimating the SI channel. Section 4 will present some solutions, that will be compared in Section 5.
4 Asynchronous Full Duplex Receiver

This section will explain the asynchronous full duplex receiver. First, the full duplex OFDM receiver is introduced, emphasizing the most important blocks for the digital interference cancellation. The channel estimation algorithms will be explained next in Section 4.2, starting from state-of-the-art OFDM channel estimation and how to extend for asynchronous full duplex.

4.1 Full Duplex OFDM Receiver

For the results, a 802.11a transceiver is programmed in LabVIEW. The 802.11a standard is chosen because its physical layer is widely used. This standard specifies an OFDM physical layer on 5GHz. Afterwards the implementation is tested using the NI USRP software defined radio platform.

Figure 5 shows a block diagram representation of the full duplex receiver implementation. The coarse timing block calculates an autocorrelation on the received signal and looks for a peak in the output. The special structure of the short preamble of the 802.11a ensures there will be a peak in the autocorrelation. The fine timing estimation calculates the cross correlation of the received signal with the known long preamble. The estimation of the $H_{2\to1}$ channel uses the FD estimation explained in section 4.2.1.

The full duplex receiver works as follows. The samples coming from the ADC are filtered. Next the receiver synchronizes on the packet of node 2, using the preamble of the packet. This synchronization is used to locate the preamble of node 2 in the stream of samples. The preamble is then used to estimate the wireless channel between node 2 and node 1. If $\delta$ in figure 4 is larger than the length of the preamble, the preamble is free of noise and the estimation will be accurate. This channel estimation will later be used for the OFDM demodulation. Using the synchronization of the preamble of node 2, the next peak in the autocorrelation can be located to find the preamble of the packet of node 1. This preamble is not free of noise as indicated in the previous section. The channel estimation algorithms will be explained in section 4.2. After estimating
the SI channel, it will be applied to the samples coming from the transmitter. These generated samples will then be subtracted from the samples coming from the receiver. Afterwards the useful signal can be demodulated.

4.2 Channel estimation

This paper compares a couple of channel estimation techniques and improves upon them for asynchronous full duplex OFDM. The following notations will be used in this paper. \( A \) represents a frequency domain value, \( a \) represents a time domain value and \( \mathbf{A} \) is a frequency domain vector. All techniques are based on the Least Squares (LS)\[9\] algorithm for its low complexity. We start with the FD channel estimation, and then propose alternatives for asynchronous full duplex OFDM.

4.2.1 Frequency domain LS and modification

FD channel estimation computes an estimate of the channel in the frequency domain, and hence calculates one tap per subcarrier. \( \mathbf{X} = (X[0], \ldots, X[N - 1]) \) is vector with the training symbols in the \( N \) sub bands of a single OFDM symbol. \( \mathbf{Y}^{(m)}, m = 1, \ldots, M \) are the received OFDM symbols with \( M \) the total number of received OFDM symbols. \( \hat{H}[k] \) is the channel per subcarrier \( k \), estimated using the LS algorithm, it is equal to

\[
\hat{H}[k] = \frac{1}{M} \left( \frac{1}{X[k]} \sum_{m=1}^{M} Y^{(m)}[k] \right).
\]

When the channel estimate is used to apply digital cancellation and not for channel equalization, an IFFT is applied to obtain the time domain channel estimate. The 802.11a preamble consists of 2 symbols, so \( M = 2 \) when using the default channel estimation on the preamble. Simulations in section 5 will show that this channel estimate is not accurate enough when interference from other transmissions is present. Because the full duplex node has knowledge of the whole transmitted package it can use it to calculate an estimate over a larger numbers of symbols, and hence average out more interference or noise. In this case \( M \) will be equal to the total number of OFDM symbols in the packet.

4.2.2 Time domain LS

The time domain LS algorithm calculates a channel estimate of \( L \) taps. \( L \) will be equal to 1 in our simulations because the SI channel has a strong line-of-sight component. The estimation uses one OFDM symbol, this means that the estimate is an average of 64 symbols. The channel estimate is by its own robust enough against noise from Node 2 because it averages out the noise. If \( \mathbf{x} = \text{FFT}(\mathbf{X}) \), a toeplitz matrix \( \mathbf{a} \) can be defined as

\[
\mathbf{a} = \begin{bmatrix}
x[0] & x[1] & \cdots & x[N - L] \\
\vdots & \ddots & \ddots & \vdots \\
x[N - L - 1] & x[N - L - 2] & \cdots & x[N - 1]
\end{bmatrix}
\]

with \( L \) the number of taps of the channel estimate. From the LS algorithm it follows that the time domain channel estimate is equal to

\[
\mathbf{h}_{SI} = (\mathbf{a}^T \mathbf{a})^{-1} \mathbf{a} \mathbf{y}.
\]

The number of entries in \( \mathbf{h}_{SI} \) is equal to the number of taps \( L \).
### 4.2.3 Applying the digital cancellation

To apply the channel, the channel estimate $\hat{h}_{SI}[n]$ is convoluted with the transmitted samples $x[n]$. The output $i[n]$ of this filter operation is equal to

$$i[n] = \sum_{k=0}^{N-1} \hat{h}_{SI}[k] x_1[n - k].$$

The output of the coarse and fine timings is used to align $i[n]$. Next $i[n]$ is subtracted from the received samples $y[n]$

$$\hat{y}[n] = y[n] - i[n]$$

$$= \sum_{k=0}^{N-1} h_{2\rightarrow1}[k] x_2[n - k] + \sum_{k=0}^{N-1} \left( h_{SI}[k] - \hat{h}_{SI}[k] \right) x_1[n - k] + z[n].$$

### 5 Results

Figure 6 shows the simulation results. The RwS line shown in the figure is simulated using the time domain LS algorithm, this is the best algorithm found in the state-of-the-art. The channel estimate is averaged over 64 time domain symbols. For the SwR scenario there are three algorithms simulated. The first one is the FD LS algorithm. Its performance is unacceptable, providing bit error rates above $10^{-2}$. This is mainly due to the bad averaging in noisy situations. Note however, that for OFDM equalization, this channel estimation algorithm is sufficient. The second, the time domain LS algorithm, provides better performance, being at most 3dB worse in BER than the RwS scenario. When using the same algorithm in both scenarios the slightly worse performance in the SwR scenario, comes from the added noise from other nodes. The modified FD LS algorithm achieves better performance across the board, approaching the performance of the RwS scenario.

![Figure 6: Simulation results](image)

Not only the BER performance is increased using the modified FD LS algorithm. The packet reception ratio (PRR) for 15dB transmit power increases from 0% using FD LS, to 58.77% using time domain LS, to 83.33% using the modified FD LS, to 96.28% for the RwS scenario. For 20dB transmit power the PRR increases from 0% to 96.80% to 98.40%, the RwS scenario’s performance is slightly worse here with 91.91%.
6 Conclusion and future work

We’ve presented a novel channel estimation technique for full duplex digital self-interference cancellation. The technique uses the full packet instead of only the preamble to estimate the channel. This method averages out more of the noise from the other node. Simulation results have shown a 2dB increase in BER performance compared to the current state-of-the-art. Achieving an almost 30% increase in PRR over the state-of-the-art.

One of the main problems with the current implementation is the synchronization performance. Because of the AGC settling in the SwR scenario, it is not possible to cancel the whole SI signal. The peak in the auto-correlation coming from the preamble of Node 1 remains and the receiver cannot synchronize on the packet from Node 2. In the RwS scenario, the problems comes from the fact that the preamble has to be found in a really noisy environment with some residual self-interference present.

References


A Computationally Efficient Soft-Output Lattice Reduction-Aided Selective Spanning Sphere Decoder for Wireless MIMO systems

Hoang Duy Nguyen†‡, Ubaid Ahmad‡, Min Li‡, Liesbet Van der Perre‡, Rudy Lauwereins‡, Sofie Pollin‡
† Brussels School of Engineering (Université Libre de Bruxelles) — Avenue F. D. Roosevelt 50, Brussels, B-1050, Belgium
‡ Interuniversity Micro-Electronics Center (IMEC) vzw — Kapeldreef 75, Leuven, B-3001, Belgium
Email: honguyen@ulb.ac.be, {ubaid, limin, vdperre, lauwerei, pollins}@imec.be

Abstract
In recent years, the algorithmic optimizations and implementations of near-optimal Multiple-Input Multiple-Output (MIMO) detectors have been an area of active research. Lattice Reduction (LR) has shown to be a promising technique to improve the performance of linear MIMO detectors. However, LR-aided linear hard-output MIMO detection is still far from optimal. Practical systems use soft-output information to exploit gains from coded systems in order to yield near-optimal performance. In this paper, the LR-aided Selective Spanning Sphere Detection algorithm is proposed as a reduced-complexity candidate list generation method for soft-output MIMO detection, specifically optimized for practical MIMO-OFDM systems. This algorithm uses efficient and scalable heuristics based on simple processor-friendly operations that significantly contribute to lowering the computational complexity of the MIMO detection problem. Results from Monte Carlo simulations reveal that LR-aided SSSD is a promising algorithm that is capable of providing near-optimal performance whilst being especially computationally efficient, in comparison to other algorithms.

Index Terms
Lattice Reduction, MIMO detection, Sphere Decoder.

I. INTRODUCTION

Multiple-Input Multiple-Output (MIMO) communication currently represents the cutting edge technology in the landscape of modern wireless communications, after theoretical investigations revealed that a linear channel capacity increase could be achieved by exploiting a multiple-antenna scheme at both sides of the wireless link, in a sufficiently rich scattering environment. Upcoming wireless standards, such as 3GPP LTE-Advanced (LTE-A) and 802.11ac for instance, all use MIMO technology in order to achieve high throughput, high spectral efficiency and improved channel reliability.

In a spatially multiplexed MIMO system, several independent data streams are transmitted simultaneously using the same physical resources, which requires the receiver to separate the received streams. For coded MIMO channels with Orthogonal Frequency-Division Multiplexing (OFDM), the Maximum A Posteriori (MAP) detector yields optimum performance with respect to Bit Error Rate (BER), at the expense of a prohibitive computational complexity: its computational complexity grows exponentially with the number of transmit antennas, making it impractical for deployment on battery-operated systems. The major challenge is thus to design a sustainable MIMO detector that achieves near-MAP performance, with a significantly reduced complexity.

Suboptimal MIMO detection schemes have been devised extensively in the literature: Sphere Decoder (SD) is widely recognized as a promising substituted approach to achieving optimal performance with reduced complexity, yet it requires a closest lattice point search within a hypersphere and its fluctuating complexity with the channel conditions is an issue. A fixed-complexity SD scheme (FSD) has been proposed in [1] to solve this drawback. Linear detectors based on the principles of zero-forcing (ZF) or
minimum mean square error (MMSE) are especially computationally efficient, but their performance is degraded under certain channel conditions.

Lattice reduction (LR) aided detectors were proposed in [2] to improve the performance of linear detectors, by performing detection using a better-conditioned channel matrix whose column vectors are nearly mutually orthogonal. Although LR-aided hard-output linear MIMO detectors are capable of achieving the same diversity order as the Maximum Likelihood (ML) detector, it was shown in [3] and [4] that LR-aided soft-output MIMO schemes can reduce the remaining performance gap with the MAP detector. Different from hard detectors which only estimate the information bits with hard decision, soft detectors also obtain reliability information about those decisions. The method proposed in [5] uses the covariance matrix of the noise along with a nearest-neighbor search method to reduce the complexity of LR-aided soft-output MIMO detection. However, the complexity of these algorithms is still very high for practical implementation. Three soft-output LR-aided MIMO detection methods were proposed in [6]. The Fixed Candidates Algorithm (FCA) and Fixed Radius Algorithm (FRA) are based on the K-best and Sphere Detection approaches, respectively. But although FCA in [6] involves modular and repetitive operations than can easily be parallelized, its complexity is too high.

In order to reduce the complexity of the detection problem, the LR-aided Selective Spanning Sphere Detection (SSSD) algorithm is proposed. It generates a list of candidate symbol vectors using an enumeration strategy derived from the Fast Enumeration (FE) algorithm in [7] and adapted for SD. Like FE, it uses efficient and scalable heuristics, enabling different performance-complexity tradeoffs to be reached.

The remainder of this paper is organized as follows: Section II describes the system model and the lattice basis reduction process. The proposed algorithm is then presented in Section III. Experimental results are reported in Section IV. Finally, conclusions are drawn in Section V.

II. LATTICE REDUCTION-AIDED MIMO DETECTION

A. Lattice Reduction-aided Hard-Output MIMO Detection

Consider a spatially multiplexed MIMO system with $N_t$ transmit and $N_r$ receive antennas undergoing an i.i.d. flat Rayleigh fading with unit variance ($N_t = N_r$ is assumed in this paper).

The vector of received symbols $y \in \mathbb{C}^{N_r \times 1}$ is given as

$$ y = Hs + n, \quad (1) $$

where $s \in \mathbb{C}^{N_t \times 1}$ denotes the vector of transmitted symbols taken independently from a $M$-ary Quadrature Amplitude Modulation constellation with $E[sH] = 1_{N_t}$, $n \in \mathbb{C}^{N_r \times 1}$ is the vector of independent complex gaussian noise samples distributed as $\mathcal{N}(0, \sigma^2)$ and $H \in \mathbb{C}^{N_r \times N_t}$ is the MIMO channel matrix, considered perfectly known at the receiver.

The MIMO detection problem consists in estimating the transmitted symbol vector $s$, from the knowledge of the received symbol vector $y$ and the channel matrix $H$.

The maximum likelihood (ML) estimate of the transmitted symbol vector is defined as

$$ \hat{s}_{ML} = \arg\min_{s \in \Omega^{N_t \times 1}} \|y - Hs\|^2, \quad (2) $$

where $\Omega^{N_t \times 1} \subset \mathbb{C}^{N_t \times 1}$ is the set containing all the possible $N_t \times 1$ complex vector symbols $s = [s_1, \ldots, s_{N_t}]$.

However, the complexity of the ML MIMO detection increases exponentially with the number of transmit antennas $N_t$, making it impractical for higher order systems.

A LR-aided linear MIMO detection scheme has been proposed in [2] to improve the performance of linear detectors. Assuming a full rank channel matrix $H$, a reduced lattice basis is first obtained as $\tilde{H} = HT$ using the Complex Lenstra-Lenstra-Lovász (CLLL) algorithm [8][9], with $T \in \mathbb{CZ}^{N_t \times N_t}$ a unimodular matrix with $|\det(T)| = 1$, where $\mathbb{CZ} = \{a + jb \mid a, b \in \mathbb{Z}\}$ denotes the gaussian integers set.
Note that the columns of $\mathbf{H}$ and $\tilde{\mathbf{H}}$ generate the same lattice since $\mathbf{T}$ is unimodular [2]. Also, the CLLL algorithm is considered optimal for LR because of its polynomial runtime.

The system equation (1) can be rewritten as
\[
  y = (\mathbf{HT})(\mathbf{T}^{-1}s) + n = \tilde{\mathbf{H}}(\mathbf{T}^{-1}s) + n
\]
\[
  \Rightarrow \tilde{\mathbf{H}}^\dagger y = \mathbf{T}^{-1}s + \mathbf{T}^{-1}\mathbf{H}^\dagger n = \mathbf{T}^{-1}\left[a\left(\bar{s} + \frac{1}{2}1_v\right)\right] + \mathbf{T}^{-1}\mathbf{H}^\dagger n,
\]
where $s = a\left(\bar{s} + \frac{1}{2}1_v\right)$, $\mathbf{T} = (\tilde{\mathbf{H}}^\dagger\tilde{\mathbf{H}})^{-1}\tilde{\mathbf{H}}^\dagger$ denotes the left Moore-Penrose pseudo-inverse of the transformed channel matrix $\tilde{\mathbf{H}}$, $\bar{s}$ are the points of the scaled and shifted QAM constellation in the $\bar{s}$-domain where all points are complex integers, and the scaling factor $a$ is $[\sqrt{2}, \sqrt{4}/10]$ in case of QPSK and 16-QAM, respectively. The shift vector $1_v$ is a $N_t \times 1$ vector of $[1 + 1_j]$. Note that the multiplication of $y$ by $\mathbf{H}^\dagger$ in (3) instead of $\tilde{\mathbf{H}}^\dagger$ causes less noise amplification since the column vectors of $\mathbf{H}$ are more orthogonal than $\tilde{\mathbf{H}}$.

Now (3) can be rewritten as
\[
  \frac{1}{a}\tilde{\mathbf{H}}^\dagger y - \frac{1}{2}\mathbf{T}^{-1}1_v = \mathbf{T}^{-1}\bar{s} + \frac{1}{a}\mathbf{T}^{-1}\mathbf{H}^\dagger n
\]
\[
  \hat{z} = \mathbf{T}^{-1}\bar{s} + w.
\]
The LR-aided hard ZF estimate is obtained by rounding $z_r = \lfloor \hat{z} \rfloor$ to the nearest gaussian integer and then transforming back to the original QAM constellation domain [2].

**B. Lattice Reduction-aided Soft-Output MIMO Detection**

The goal of soft-output MIMO detection is to obtain reliability information about the detected symbols. It usually consists of two parts: (a) A list generator that gives a list of candidate symbol vectors, denoted by $\mathcal{L} \subseteq \Omega^{N_t}$; (b) A Log-Likelihood-Ratio (LLR) generator that approximates the A Posteriori Probabilities (APP) for each bit; the approximation becomes near optimal when $\mathcal{L} = \Omega^{N_t}$.

Generating the candidate list dominates the performance and the complexity of the soft-output detection; if the candidate list is too large the performance is near-optimal but the complexity is too high, and if the list size is too small the performance is close to hard MIMO detection.

Sphere Decoding (SD) solves the ML detection problem by considering the QR-decomposition of the channel matrix as $\mathbf{H} = \mathbf{Q}\mathbf{R}$ where $\mathbf{Q}$ is a unitary matrix ($\mathbf{Q}^\mathbf{H}\mathbf{Q} = \mathbf{I}$) and $\mathbf{R}$ is an upper triangular matrix, so as to replace (2) by the equivalent expression
\[
  \hat{s}_{SD} = \arg\min_{\bar{s} \in \Omega^{N_t}} \left[\|\hat{y} - \mathbf{R}\bar{s}\|^2 + c\right], \quad \left\{\begin{array}{l}
  \hat{y} = \frac{1}{a}\mathbf{Q}^\mathbf{H}y - \frac{1}{2}\mathbf{R}1_v \\
  s = a\left(\bar{s} + \frac{1}{2}1_v\right)
  \end{array}\right.
\]
where $c$ is a constant.

Rather than a depth-first search like many SD implementations found in the literature, here a breadth-first spanning-tree is constructed to generate a set of candidates minimizing the distance in (6). The level of the tree is $N_t + 1$; mark the root-level as $i = N_t + 1$ and the leaf-level as $i = 1$. Each node at level $i \in \{N_t + 1, \ldots, 2\}$ is expanded to $\mathcal{M}$ nodes at level $i + 1$, where $\mathcal{M}$ is the constellation size. In this tree each node at level $i \in \{N_t, \ldots, 2, 1\}$ is uniquely described by the partial vector symbols $\bar{s}^i = [\bar{s}_i, \bar{s}_{i+1}, \ldots, \bar{s}_{N_t}]$, the leaves at level $i = 1$ correspond to all possible $N_t \times 1$ vector symbols. Annotate the root node with $T_{N_t+1} = 0$ and starting from Level $i = N_t$, the PED (Partial Euclidean Distance) of partial symbol vector $\bar{s}^i = [\bar{s}_i, \bar{s}_{i+1}, \ldots, \bar{s}_{N_t}]$ is $T_i(\bar{s}^i) = T_{i+1}(\bar{s}^{i+1}) + \|e_i(\bar{s}^i)\|^2$, where the PED-increment is $\|e_i(\bar{s}^i)\|^2 = \|\hat{y}_i - \sum_{j=1}^{N_t} R_{ij}\bar{s}_j\|^2$, where $R_{ij}$ are the $i$th row and $j$th column entries of $\mathbf{R}$. Since $\|e_i(\bar{s}^i)\|^2$ is non-negative, the PED increases monotonically from root to leaves. The optimal solution then consists in finding the leaf at level $i = 1$ with the minimal PED, $T_1(\bar{s}^1)$. 
III. LATTICE REDUCTION-AIDED SELECTIVE SPANNING SPHERE DETECTOR

The runtime-adaptive LR-aided SSSD detector is proposed to reduce the complexity of spanning-sorting-deleting operations inherent to the breadth-first tree-search in the K-best and K-best sphere decoding algorithm [10]. The idea is to first generate the LR-aided ZF hard estimate $\hat{s}_{LR-ZF}$ in the $\bar{s}$-domain to determine the extent of the radius search, and then a candidate list is built by enumerating efficiently constellation points inside this radius using low-complexity heuristics.

Simulation results reveal that these characteristics significantly contribute to reducing the complexity of the detection problem, with limited performance degradation.

A. Selective Spanning

The SSSD algorithm determines the list of tree leaves that satisfy the inequality

$$\left\| \hat{y}_i - \sum_{j=i}^{Nt} R_{ij} \bar{s}_j \right\|^2 \leq r^2, \quad r^2 = \| \hat{y} - R\bar{s}_{LR-ZF} \|^2 \quad (7)$$

where $r$ is the radius constraint.

This algorithm is uniquely characterized by a $Nt \times 1$ vector $m = [m_1, \ldots, m_{Nt}]$. Starting from root level $i = Nt$, SSSD spans each node at level $i + 1$ to at most $m_i$ nodes at level $i$. If the node at level $i = Nt + 1$ has the associated partial symbol vector being $\bar{s}^{i+1} = [\bar{s}_{i+1}, \ldots, \bar{s}_{Nt}]$, the spanning is to select a set of $\bar{s}^i = [\bar{s}_i, \bar{s}_{i+1}, \ldots, \bar{s}_{Nt}]$ in a way that the PED-increment satisfies the radius constraint in (7). Hence, depending on the channel conditions, the total number of nodes at leaf level is at most $\prod_{k=1}^{Nt} m_k$.

The topologies of breadth-first K-best and SSSD search trees are compared on Fig. 1, where $K = 4$ for K-best and $m = [1 \ 2 \ 2 \ 4]$ for SSSD, for a possible detection. The unspanned nodes for SSSD are represented by the dashed lines in Fig.1(b).

At the contrary of K-best which systematically searches for exactly K-best candidates at each layer and has a fixed complexity, SSSD instead aims to find at most $m_i$ candidates at layer $i$ that satisfy best the radius constraint in (7).

As a result, SSSD has a complexity that evolves dynamically with the channel conditions and the noise level; improved channel conditions will bring a shrinkage of the search radius $r$, resulting in a decrease in the complexity of the detection. What is more, the overall complexity of SSSD is upper bounded, as explained below.

B. Fast Sphere Enumeration

The idea of the Fast Sphere Enumeration (FSE) algorithm is to provide efficient, low-complexity heuristics that approximate the sorting-deleting operations inherent to K-best.
To derive this algorithm, (7) is rewritten as follows:

\[
\begin{aligned}
\|\hat{y}_i - \sum_{j=i}^{N_t} R_{ij}\bar{s}_j\|^2 &= \left\|\hat{y}_i - \sum_{j=i+1}^{N_t} R_{ij}\bar{s}_j - R_{ii}\bar{s}_i\right\|^2 < r^2 \\
\|\hat{y}_k - \sum_{j=i+1}^{N_t} R_{ij}\bar{s}_j - R_{ii}\bar{s}_i\|^2 < \frac{r^2}{R_{ii}^2}
\end{aligned}
\] (8)

This can be done, since minimizing \(\|e_i (s^i)\|^2\) is equivalent to minimizing \(\|e_i (\bar{s}^i) / R_{ii}\|^2\).

Geometrically speaking, equation (8) consists in finding at layer \(i\) a set of at most \(m_i\) constellation points closest to \(\xi_i := (\hat{y}_i - \sum_{j=i+1}^{N_t} R_{ij}\bar{s}_j) / R_{ii}\) in the \(\bar{s}\)-domain and enclosed inside a circle of radius \(r' := r / R_{ii}\).

This problem can be approximated to reduce the complexity of the detection. Let \(\bar{x}\) be the set of points that lie inside the aforementioned circle and also within the \(\bar{s}\)-domain boundaries. The problem (8) can then be rewritten as follows:

\[
\bar{x} = \{x_i \in \bar{s}\text{-domain} \subset \mathbb{C}\mathbb{Z}\} \left\{ \begin{array}{l}
\Re(x_i) \in [x_{\min}, x_{\max}] \\
\Im(x_i) \in [y_{\min}, y_{\max}]
\end{array} \right.
\] (9)

where

\[
\begin{aligned}
x_{\min} &= \max \left\{ \Re(\xi_i) - r', -\sqrt{M/4} \right\} \\
x_{\max} &= \min \left\{ \Re(\xi_i) + r', \sqrt{M/4} - 1 \right\} \\
y_{\min} &= \max \left\{ \Im(\xi_i) - r', -\sqrt{M/4} \right\} \\
y_{\max} &= \min \left\{ \Im(\xi_i) + r', \sqrt{M/4} - 1 \right\}.
\end{aligned}
\]

Note that \(-\sqrt{M/4}\) and \(\sqrt{M/4} - 1\) are simply the boundaries of the constellation in the \(\bar{s}\)-domain (e.g. for a 64-QAM constellation, the set of points are \(\{-4, \ldots, 3\} + j \{-4, \ldots, 3\}\)).

Concretely, the detection problem can now be seen as the enumeration in ascending Euclidean distance of at most \(m_i\) points closest to \(\xi_i\) inside the rectangular \(\bar{x}\)-domain.

In order to do so without resorting to inefficient sorting operations that require much memory rearrangement, the vectorized and scalable FSE algorithm (Alg. 1) provides an efficient enumeration strategy that is able to handle any QAM constellation without compromising the algorithm dynamism and with very little complexity-increment; the enumeration strategy is based on the Fast Enumeration (FE) algorithm in [7], which implies that only simple arithmetics are used, such as bitwise binary addition, subtraction, bit-not, shift and slicing operators, which can all be implemented efficiently with low hardware cost.

The FSE strategy is as computationally efficient as the Fincke-Pohst or Schnorr-Euchner enumeration strategies [11].

For instance, the first 9 points are enumerated as follows, using the same notations as in Alg. 1:

\[
p_{1\rightarrow 9} = Q(\xi_i) + \text{sign}(\Re(d)) \cdot A + j \cdot \text{sign}(\Im(d)) \cdot B
\] (10)

with \(Q(\cdot)\) being the slicing operator (rounding and boundary check) and \(A = [0, \phi, \bar{\phi}, 1, 0, 1, -1, -1, -1]^T\) and \(B = [0, \phi, \bar{\phi}, 1, -1, 0, -1, 0, 1, -1]^T\).

If more points are required, the same rule of thumb is used to incrementally grow the set \(\bar{x}\) around \(\xi_i\) using \(d\) and \(\phi\) as guidelines to direct the order of the enumeration.

For the sake of the illustration, Fig. 2 shows the enumeration for a given radius for a 64-QAM constellation in the \(\bar{s}\)-domain, with \(m_i = 8\). Since there are only 6 points inside the bounding rectangle, only 6 constellation points will be returned.
Algorithm 1: Fast Sphere Enumeration

**Input:** \(\xi_i, r'_i, m_i\)

1. Compute \(d \leftarrow \xi_i - Q(\xi_i)\);
2. Compute \(\phi \leftarrow |\Re(d)| > |\Im(d)| \& \overline{\phi} = \text{NOT}(\phi)\);
3. Compute \(\Re\) extension \(a \leftarrow \lceil \Re(\xi_i) + r'_i \rceil - \lfloor \Re(\xi_i) - r'_i \rfloor\);
4. Compute \(\Im\) extension \(b \leftarrow \lceil \Im(\xi_i) + r'_i \rceil - \lfloor \Im(\xi_i) + r'_i \rfloor\);
5. Compute \(x_{\min}, x_{\max}, y_{\min}, y_{\max}\);
6. For \(k = 1\) to \(\lceil \max(a, b) / 2 \rceil\) do
   - \(x \leftarrow \text{Use modified FE algorithm to build a list of constellation points inside a square of size } k\);
   - \(\tilde{x} \leftarrow \text{Keep up to } m_i \text{ elements in } x \text{ that verify } \Re(x_j) \in [x_{\min}, x_{\max}] \& \Im(x_j) \in [y_{\min}, y_{\max}]\)
7. Output: \(\tilde{x}\)

IV. EXPERIMENTAL RESULTS

In this section, the BER performance and average complexities of the following detection algorithms are assessed, considering a 4 x 4 1/2 turbo-coded LTE channel with complete channel state information at the receiver:

- Zero-Forcing (ZF)
- LR-aided Zero-Forcing (CLLL-ZF)
- Breadth-first LR-aided FCA/FRA/FMA [6]
- Breadth-first LR-aided SSSD
- Maximum Likelihood (ML)
- Maximum A Posteriori (MAP)

A. BER Performance

For the sake of limited place, herein only a 64-QAM channel is shown, on Fig. 3. As expected, all detectors perform significantly better than ZF and CLLL-ZF. SSSD with only one candidate \(m = [1 \ 1 \ 1 \ 1]\) provides performance gain compared to CLLL-ZF, and is only 1.2 dB away from ML, for a BER of \(10^{-4}\). With \(m = [1 \ 1 \ 1 \ 32]\), we are only 1.6 dB away from MAP. This gap can be further reduced by using bigger values of \(m_i\).

B. Complexity Estimation

The average computational complexities of the simulated detectors are compared on Table I.
The estimated complexities were obtained by counting the average (equivalent) number of mathematical operations required to detect one MIMO symbol in a $4 \times 4$ QAM channel, assuming the following cost model for the operations: (a) real operations: 1, (b) complex additions: 2, (c) complex multiplications: 6, (d) sorting operations: $O(n \log n)$.

Preprocessing cost (QR-decomposition, lattice reduction using the CLLL algorithm, matrix inversions) is also included, and accounts for about 1800 additions and 500 multiplications for LR-aided detectors on average.

By analyzing Table I, we assess the benefits of the SSSD algorithm. For instance, SSSD with $m = [1 1 1 32]$ provides almost the same BER performance as FCA with $K_p = 8$ (FCA being slightly better by 0.5 dB at $10^{-4}$), but the number of required operations is significantly reduced (by a factor of 9, in Table I).

Although the size of the candidate list of the former is bigger than the latter to achieve similar BER performance, the cost of generating the candidate list using FSE is significantly lower, since there is no need to sort symbols in ascending Euclidean distance.

Moreover, changing $m$ has a direct impact on the overall quality/cost of the detection. This unique feature implies that SSSD along with the FSE algorithm as list generator is a scalable and efficient algorithm that can easily be tuned to meet given complexity-performance requirements. Also, the $m$ vector can be adjusted to respond dynamically to varying channel conditions. This is especially desirable for emerging software-defined radios (SDR) baseband architectures.

V. CONCLUSION

In this work, the LR-aided Selective Spanning Sphere Detection algorithm was proposed as a reduced-complexity soft-output MIMO detector that achieves an attractive performance-complexity compromise,
parametrized by the \( m \) vector. Simulation results revealed that under equivalent performance constraints, the SSSD detector clearly outperformed other existing alternatives in terms of computational complexity without jeopardizing the performance.

VI. RELATION TO PRIOR WORK

The LR-aided SSSD algorithm utilizes the selective spanning tree-search strategy introduced in [7] and combines it with sphere detection [12]. The work presented in [6] on LR-aided MIMO detection takes a fixed sphere radius for SD, whereas this paper shows that a sphere search radius based on LR-aided ZF estimate can significantly reduce the computational complexity.

REFERENCES

Closed-loop Adaptation of a Nonlinear Interference Suppressor for Local Interference in Multimode Transceivers

H. Habibi, E. J. G. Janssen, Wu Yan, P.G.M. Baltus, J. W. M. Bergmans
Eindhoven University of Technology, Department of Electrical Engineering,
Den Dolech 2, 5612 AZ Eindhoven, The Netherlands
Email: H.Habibi@tue.nl, HabibiHooman@gmail.com, Phone: +31402473287

Abstract

In multimode transceivers, the transmitter for one communication standard induces a large interference in the receiver for another standard, exceeding the desired signal by many orders of magnitude. To linearly suppress this interference, the receiver should have a very large linear dynamic range, resulting in excessive power consumption. An adaptive memoryless nonlinearity, which requires an adaptation signal proportional to the envelope of the received interference, can be used to strongly suppress the interference without excessive power consumption. In this paper, we propose to digitally generate the adaptation signal using a model, which describes the adaptation signal in terms of the locally available baseband interference. The model is adapted during the transceiver operation such that the power of the residual interference at the output of the nonlinearity is minimized. Simulation results show that the proposed adaptation method can strongly suppress the interference while a symbol error rate close to that of an exactly linear receiver is achieved.

Index Terms

Multimode transceivers, Interference suppression, Nonlinear systems, Adaptive filters, TX leakage, co-located transceivers.

I. INTRODUCTION

Nowadays, many handheld devices have become multimode transceivers, supporting a multitude of communications standards. From the users’ point of view, the simultaneous operation of these transceivers is highly desirable. However, due to the small size of the handheld device, the Local Transmitter (LTX) of one standard induces a strong interference in the Local Receiver (LRX) of another standard [1]. To suppress this local interference by linear filtering the receiver should have a very large linear dynamic range, resulting in excessive power consumption [2].

An alternative approach to linear filtering is to suppress the interference by passing the received signal through an adaptive Nonlinear Interference Suppressor (NIS) [2] [3]. The Input-Output (IO) characteristic of the NIS can be modeled as the combination of a hard limiter IO with an adaptable limiting amplitude $l(t)$ and a linear IO (with gain of $-c$), as shown in Fig. 1. In [2], it is shown that for an interference with an envelope $A_i(t)$ at the NIS input, there is an optimal adaptation signal:

$$\tilde{l}(t) = \frac{\pi}{4} c A_i(t),$$

which by adapting the NIS according to it the following goals are achieved:
Goal 1: Suppress the interference such that the power of the interference will be smaller than power of the desired signal at the NIS output.

Goal 2: Introduce a negligible amount of nonlinear distortion to the desired signal.

In [2], the NIS approach is studied with the assumption that the optimal adaptation signal is known. To calculate the optimal adaptation signal according to (1), c and $A_i(t)$ must be known. In the multimode transceiver a baseband version of the transmitted interference is locally available. By identifying a baseband model of the coupling path of the interference from the transmitted baseband interference to the received interference at the NIS input, $A_i(t)$ can be estimated. The coupling path is subject to environmental changes, e.g. the presence of the user’s hand. Hence the path model must be continuously adapted during the transceiver’s operation.

In this paper, we develop a closed-loop method to adapt the path model such that the power of residual interference at the NIS output is minimized. Simulation results show that the proposed adaptation method can strongly suppress the interference while a symbol error rate close to that of an exactly linear receiver is achieved.

II. SYSTEM MODEL

In this section we describe the model of the multimode transceiver that uses the NIS. This model will be used to analyze the effect of the NIS on the receiver operation and estimation of the adaptation signal.

A. Description of the signals received by the local RX

The model shown in Fig. 2 includes the LTX and the LRX. At the LRX, a desired signal transmitted by the remote TX is received in the presence of a part of the transmitted interference coupled from the LTX. The combination of these two signals is passed through a Band Pass Filter (BPF1). Typically a SAW filter is used for BPF1. The desired signal is passed essentially unchanged through BPF1 and the interference is attenuated to some
extent by BPF1. After BPF1, the NIS input \( x(t) \) includes both a desired signal \( x_d(t) \) and an interference \( x_i(t) \) as:

\[
x(t) = x_d(t) + x_i(t) = A_d(t) \cos(2\pi f_d t + \varphi_d(t)) + A_i(t) \cos(2\pi f_i t + \varphi_i(t)),
\]

where \( A_d, \varphi_d, f_d, A_i, \varphi_i, \) and \( f_i \) are envelope, phase and center frequencies of the desired signal and interference after BPF1, respectively. The desired signal is bandlimited to \([f_d - B_d/2, f_d + B_d/2]\) and the interference is bandlimited to \([f_i - B_i/2, f_i + B_i/2]\), where \( B_d \) and \( B_i \) are bandwidths of the desired signal and interference, respectively. After BPF1, \( x(t) \) is passed through the NIS which is adapted by an adaptation signal \( l(t) \). Average SIR at the NIS input is defined as: \( \text{SIR}_s = \frac{E(A_d^2)}{E(A_i^2)} \), where \( E() \) denotes statistical expectation. Since the NIS is a strongly nonlinear circuit, high frequency harmonics (at frequencies around \( 3f_i, 5f_i, \) etc) are also generated at the NIS output. These harmonic are far from \( f_d \) and they are filtered out with a simple band pass filter (BPF2).

### B. Description of the Adaptation signal

In this section we present a model that describes the required adaptation signal in terms of the baseband interference which is locally available. As shown in Fig. 2, the complex-valued baseband interference \( i[p] \) with a baud rate \( \frac{1}{T_i} \) is up-sampled by an integer factor \( r_i \) by inserting zeros between samples of \( i[p] \) (complex-valued signals are shown with solid bold lines). The up-sampled signal \( i[n] \) is passed through a pulse shaping filter, resulting in a signal \( i_s[n] \). A Digital to Analog Converter (DAC) with a conversion period of \( T = \frac{T_i}{r_i} \) converts \( i_s[n] \) to an analog baseband signal \( i_b(t) \), and the TX Front-End (FE) up-converts \( i_b(t) \) to a center frequency \( f_i \). A part of the transmitted signal \( i_t(t) \) is coupled to the LRX and after passing through BPF1 is received at the NIS input.

In Fig. 2, the coupling path of the interference from \( i[n] \) to \( x_i(t) \) is shown with a dashed bold line. This path can be modeled as a linear system with a complex-valued baseband impulse response \( h(t) \). Hence the optimal adaptation signal \( l(t) \) is obtained as:

\[
\tilde{l}(t) = \frac{\pi}{4} c A_i(t) = \frac{\pi}{4} c |x_i(t)| = \left| \sum_{m=-\infty}^{\infty} i[m]h(t - mT) \right|.
\]

In (3), the scaling factor \( \frac{\pi}{4} c \) is considered as part of \( h(t) \). To digitally generate \( l(t) \) a discrete-time representation of \( l(t) \) is required. Using the following notations for signals and impulse responses at time \( t = nT \):

\[
\tilde{l}[n] = \tilde{l}(nT), \quad A_i[n] = A_i(nT), \quad h_n = h(nT),
\]

and considering the causality of \( h_n \), we can obtain the discrete-time counterpart of (3) as:

\[
\tilde{l}[n] = \frac{\pi}{4} c A_i[n] = |(h \ast i)[n]| = \left| \sum_{m=0}^{+\infty} i[n - m]h_m \right|.
\]

Here we assume that the sampling frequency \( \frac{1}{T} \) is high enough so that \( \tilde{l}(t) \) can be reconstructed from \( \tilde{l}[n] \) with a negligible error. Our goal here is to determine a set of filter taps \( g_n \) such that the power of the residual interference at the NIS output would be minimized. These taps result in an estimate \( \hat{l}[n] \) of the adaptation signal as:

\[
\hat{l}[n] = |(g \ast i)[n]| = \left| \sum_{m=0}^{M-1} i[n - m]g_m \right|.
\]
where \( M \) taps are used to realize \( g \). A DAC converts \( \hat{l}[n] \) to a continuous-time signal \( \hat{l}(t) \) which is applied as the estimated adaptation signal to the NIS.

C. Description of the signals at the NIS output

As shown in Fig. 1, the NIS output is the combination of the limiter and linear amplifier outputs. Using the approximations for the bandpass limiter output [4] for \( A_i \gg A_d \), we obtain:

\[
y(t) \simeq A_{d,y}(t) \cos(2\pi f_d t + \varphi_d(t)) + A_{i,y}(t) \cos(2\pi f_i t + \varphi_i(t)) + A_{IM}(t) \cos(2\pi (2f_i - f_d) t + 2\varphi_i(t) - \varphi_d(t)),
\]

where \( A_{d,y} \), \( A_{i,y} \) and \( A_{IM} \) are envelopes of desired signal, interference and main Inter-Modulation (IM) components at the NIS output, respectively. For \( A_i \gg A_d \) these envelopes can be approximated by [4]:

\[
A_{d,y}(t) \simeq \left( \frac{2l(t)}{\pi A_i(t)} - c \right) A_d(t) = \left( \frac{l(t)}{2l(t)} - 1 \right) cA_d(t),
\]

\[
A_{i,y}(t) \simeq \frac{4l(t)}{\pi} - cA_i(t) = \frac{4}{\pi}(l(t) - \bar{l}(t)),
\]

\[
A_{IM}(t) \simeq - \frac{2A_d(t)}{\pi A_i(t)} \bar{l}(t) = - \frac{l(t)}{2l(t)} cA_d(t).
\]

For \( l(t) = \bar{l}(t) \), it is obtained \( A_{d,y}(t) \simeq - \frac{c}{2} A_d(t) \), \( A_{i,y}(t) \simeq 0 \), and \( A_{IM}(t) \simeq - \frac{c}{2} A_d(t) \). During the receiver operation \( \bar{l}(t) \) is used as the NIS adaptation signal.

III. Closed-loop adaptation of the NIS

![Diagram of NIS adaptation](image)

Fig. 3: NIS adaptation.

Since \( h \) depends on the changes in the environmental, \( g \) must be adapted to track these changes. To this end we measure the envelope \( A_{i,y} \) of the residual interference at the NIS output and adapts \( g \) such that \( E(A_{i,y}^2) \) is minimized. As shown in Fig. 3, \( A_{i,y} \) is measured using a simple switching mixer, as will be explained in Section III-A, and is sampled by
an ADC. To adapt \( g[n] \) we process \( \eta[n] \) and \( i[n] \) together such that \( \text{E}(A_{i,y}^2) \) is minimized, as will be explained in Section III-B. Using \( i[n] \) instead of \( i_s[n] \) as the adaptation reference has two advantages. Firstly, \( i[n] \) is a white signal. Hence the adaptation converges with a single mode of convergence. Secondly, \( i[n] \) is quantized with fewer bits compared to \( i_s[n] \) and \( \frac{\nu_{i}}{\nu_{\eta}} \) of its samples are zero. This simplify the adaptation, computationally. Finally, \( \dot{l}[n] = [(g * i)[n]] \) is calculated and converted to \( \dot{l}(t) \) using a DAC.

A. Extraction of error signal

As shown in Fig. 3, to measure \( A_{i,y} \) we propose to down-convert \( y(t) \) using a switching mixer with \( x(t) \) as its Local Oscillator (LO) port and \( y(t) \) as its Radio Frequency (RF) port. A switching mixer changes the sign of its RF input based on its LO input as:

\[
\eta(t) = \begin{cases} 
    y(t) & \text{for } x(t) > 0, \\
    0 & \text{for } x(t) = 0, \\
    -y(t) & \text{for } x(t) < 0.
\end{cases}
\]  

(10)

In Appendix I, we prove that for \( A_d << A_i \):

\[
\eta(t) \approx \frac{2}{\pi} A_{i,y}(t) + \nu(t) \approx \frac{8}{\pi^2} (\tilde{l}(t) - \tilde{l}(t)) + \nu(t) = \frac{8}{\pi^2} (|g(t) * i(t)| - \tilde{l}(t)) + \nu(t)  
\]

(11)

where \( \nu(t) \) acts as a disturbance term in the estimation of \( g \).

B. adaptation algorithm

To minimize \( \text{E}(\eta^2(t)) \) we sample \( \eta(t) \) as:

\[
\eta[n] = \frac{8}{\pi^2} (|((g * i)[n]) - \tilde{l}[n]) + \nu[n] = \frac{8}{\pi^2} (|g^T i[n] - \tilde{l}[n]) + \nu[n]
\]

(12)

where column vectors \( g, i[n] \) and \( \nu[n] \) are defined as:

\[
g = [g_0, ..., g_{M-1}]^T, \quad i[n] = [i[n], ..., i[n-M+1]]^T, \quad \nu[n] = [\nu[n], ..., \nu[n-M+1]]^T,
\]

(13)

and the superscript \( T \) denotes the transpose operation. The loop adapts the filter taps \( g_n \) to minimize a cost function defined as:

\[
J(g) = \text{E} \left( \left( \frac{\pi^2}{8} \eta(t) \right)^2 \right) = \text{E} \left( (|g^T i[n]| - \tilde{l}[n])^2 \right) + \text{E} \left( \frac{\pi^4}{64} \nu^2[n] \right)
\]

(14)

The steepest decent algorithm can be used to minimize \( J(g) \) [5]. To use this algorithm the complex-valued gradient vector \( \nabla g J(g) \) of the cost function is required which can be obtained as:

\[
\nabla g J(g) = 2\text{E} \left\{ \eta[n] \frac{g^T i[n]}{|g^T i[n]|} i^*[n] \right\}
\]

(15)

where \( ^* \) denotes complex conjugate. The derivation is omitted here due to the space limitations. Approximating the expected value in (15) by its instantaneous value results in the stochastic version of the steepest decent algorithm as:

\[
g[n+1] = g[n] - \mu \eta[n] \frac{g^T[n] i[n]}{|g^T[n] i[n]|} i^*[n],
\]

(16)

where \( g[n] \) denotes the filter taps at time instant \( n \) and \( \mu \) is a positive real number called step-size.
C. Convergence of the adaptation algorithms

Generally the presence of local minima in a cost function disrupts the convergence of the steepest decent algorithm to its global minima. One can obtain the second derivative (also called Hessian) of $J(g[n])$ as:

$$
\nabla^2_g(J(g[n])) = 2E \left( \left( 2 - \frac{\tilde{l}[n]}{\bar{l}[n]} \right) i[n]i[H[n]] \right).
$$

(17)

According to (17) the cost function is not convex and the second derivative becomes zero when $\tilde{l}[n] = \frac{1}{2}\bar{l}[n]$. The cost function for $M = 1$ and $h_0 = 0.5 + 0.5i$ is shown in Fig.4. The x axis and the y axis show the real and imaginary parts of $g_0$, and z axis shows $J(g_0)$. Although $J(g)$ is not convex there is no local minimum. The global minimum occurs for all the points that have the same amplitude as $h_0$. There is one local maximum at $g_0 = 0$. Since there is no local minimum the adaptation algorithm converges to the global minimum.

IV. SIMULATION RESULTS

A. Simulation setup

We consider a multimode scenario of a WLAN LRX with a WiMAX LTX with $f_d=2460$ MHz and $f_i = 2510$ MHz. These center frequencies result in frequency separations of $\Delta f = 50$ MHz. Both signals have OFDM modulation with 64 subcarriers, where each subcarrier is modulated with 16 QAM. Baudrates of the interference and the desired signal are 20 MSPS and 10 MSPS, respectively. Root raised cosine pulse shaping with a roll-off-factor of 0.5 is used for both signals. The power of transmitted WiMAX signal is assumed 20 dBm and power of the WLAN signal can be as low as -70 dBm. We assume that there is -10 dB coupling between the LTX and the LRX. The BPF1 filter suppresses the WiMAX interference at $f_i$ by 10 dB. Hence Signal to Interference Ratio (SIR) at the NIS input can be as low as -70 dB. We assume that the WLAN signal is passed through an Additive White Gaussian Noise (AWGN) channel. Hence the SER performance of a linear receiver depends only on the desired Signal to Noise power Ratio (SNR). The SNR is chosen as 17.6 dB which results in an un-coded SER of $10^{-3}$ for 16 QAM modulation with an exactly linear RX [6]. In all simulations $c = 1$.

B. Interference suppression

Fig. 5a shows the PSD of $x(t)$. The X-axis shows the frequency in MHz with reference to $f_i$. In Fig. 5a, the interference is centered at zero frequency and the desired signal is
centered at $f_d - f_i = -50$ MHz. SIR$_x$ is -60 dB in this simulation. The input channel noise is filtered by the BPF1 filter and is centered at about -50 MHz. The BPF1 is assumed to be a SAW filter. The impact of the BPF1 on the interference is seen in Fig. 5a. Fig. 5b shows the PSD of $y(t)$ after reaching the steady state condition. We see that the interference at zero frequency is suppressed below the noise floor. The IM component is present at +50 MHz, ($2f_i - f_d$), with the same power as that of the desired signal.

C. Symbol error rate

Fig. 6 shows the un-coded SER vs. SIR$_x$ for the ideal adaptation signal based on (1) as well as the closed-loop adaptation method. For the closed-loop method the SER is measured after reaching the steady state. The adaptation is performed with two values of $\mu$ (0.0001 and 0.0003) which are equivalent to 160 Hz and 480 Hz 3-dB bandwidth of the adaptation loop, respectively. In all cases when SIR decreases SER decreases and reaches a constant level. For the ideal adaptation the SER degradation is only because of Gain Variation Distortion (GVD) and the IM leakage [2]. The SER degradation because of GVD becomes negligible for SIR$_x < -30$ dB. For the closed-loop method the disturbance component $\nu(t)$ in (11), causes a random adaptation error $\hat{l}(t) - \tilde{l}(t)$ which its power increases when $\mu$ increases. The adaptation error slightly degrades the SER compared to the ideal adaptation. However for a sufficiently small $\mu$ which still affords a practical adaptation speed, this SER degradation is negligible.

![Fig. 5: PSD of NIS input signal $x(t)$ and output signal $y(t)$](image)

![Fig. 6: SER vs. SIR$_x$ for OFDM modulation, SER of the baseline RX: $10^{-3}$](image)

V. CONCLUSION

In multimode transceivers, the interference induced by a local transmitter can be several orders of magnitude larger than the received desired signal, even after partial suppression by analog filters. Hence a linear receiver requires an excessive linear dynamic range to process
the desired signal in the presence of such a large interference, leading to an unreasonable power consumption. A Nonlinear Interference Suppressor (NIS) which is adapted to track the envelope of the received interference can suppress the interference without excessive power consumption. In this paper we propose to generate the adaptation signal using an adaptive baseband model of the coupling path of the interference. This model is adapted during the transceiver operation such that the power of the residual interference at the output of the NIS is minimized. The simulations for a practical scenario shows that the proposed method can suppress the interference to a level below that of the desired signal while introducing negligible degradation to symbol error rate of the receiver.

APPENDIX I: DERIVATION OF THE ERROR SIGNAL

In this appendix the extraction of the error signal for the adaptation loop is analyzed. The switching mixing described in (10) is equivalent to multiplying $y(t)$ by $x_L(t) = \text{sign}(x(t)) = \{1, x(t) > 0; 0, x(t) = 0; -1, x(t) < 0\}$. Equivalently $x_L(t)$ can be obtained by passing $x(t)$ through a hard limiter. Using the analysis in [4] for a bandpass limiter when $A_i(t) \gg A_d(t)$, $x_L(t)$ is obtained as:

$$x_L(t) \approx \frac{4}{\pi} \left( \cos(2\pi f_i t + \varphi_i(t)) + \frac{A_d(t)}{2A_i(t)} \cos(2\pi f_d t + \varphi_d(t)) - \frac{A_d(t)}{2A_i(t)} \cos(2\pi(f_i - f_d)t + 2\varphi(t) - \varphi_d(t)) \right)$$

$$+ \text{high frequency components around } 3f_d, 3f_i, 5f_d, 5f_i, ... 2f_i \pm f_i,...$$

(18)

The BPF2 in Fig. 3 filters out the high frequency components of $y(t)$ in (7) such that only the components around $f_d$ and $f_i$ remain. Thus it is sufficient to only consider the component of $x_L(t)$ around $f_i$, $f_d$, and $2f_i - f_d$. The mixer output $\eta(t)$ is obtained as:

$$\eta(t) \approx x_L(t) y(t) = \frac{2}{\pi} \left( A_{i,y}(t) + (A_{d,y}(t) - A_{IM}(t)) \frac{A_d(t)}{2A_i(t)} \right)$$

$$+ (A_{d,y}(t) + A_{i,y}(t)) \frac{A_d(t)}{2A_i(t)} \cos(2\pi(f_d - f_i)t + \varphi_d(t) - \varphi_i(t))$$

$$+ \text{high frequency components around } 2f_d, 2f_i, 2f_i - 2f_d, 2(2f_i - f_d), f_i + f_d,...$$

(19)

Because of the low-pass nature of the feedback loop we can neglect the high frequency components of $\eta(t)$. Also the component at $f_d - f_i$ that has an envelope of $A_{i,y}(t) \frac{A_d(t)}{2A_i(t)}$ which is much smaller than $A_{d,y}(t)$ can be neglected.

$$\eta(t) = \frac{2}{\pi} \left( A_{i,y}(t) + A_{d,y}(t) \frac{A_d(t)}{2A_i(t)} - A_{IM}(t) \frac{A_d(t)}{2A_i(t)} + A_{d,y}(t) \cos(2\pi(f_d - f_i)t + \varphi_d(t) - \varphi_i(t)) \right).$$

(20)

Using (8), (replacing $l(t)$ with the estimate adaptation signal $\hat{l}(t)$), and (20), $\eta(t)$ is simplified to:

$$\eta(t) \approx K(\hat{l}(t) - \bar{l}(t)) + \nu(t) = K(|g(t) * i(t)| - |h(t) * i(t)|) + \nu(t),$$

(21)

where

$$K = \frac{8}{\pi^2} \left( 1 + \frac{1}{2A_i^2(t)} \right) \approx \frac{8}{\pi^2}, \ \text{and} \ \nu(t) \approx \frac{1}{\pi c} A_d(t) \cos(2\pi(f_d - f_i)t + \varphi_d(t) - \varphi_i(t)$$

(22)

REFERENCES


Joint precoding vector and modulation and coding scheme recalculation for LTE-A multi-user MIMO

Hsin-Hung Chen, Rodolfo Torrea-Duran, Paschalis Tsiavflakis, Marc Moonen
Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven, Belgium
{Rodolfo.TorreaDuran, Paschalis.Tsiavflakis, Marc.Moonen}@esat.kuleuven.be

Abstract—Multi-user MIMO techniques were born due to the urge of high data rates and spectral efficiency in 4G systems. For scenarios with a large number of users to be served in one cell, high capacity gains can be achieved by transmitting independent data streams to different users sharing the same time-frequency resources through the use of MIMO precoding. With enough channel state information (CSI) at the transmitter, MIMO precoding allows to increase multi-user diversity gain. However, without a correct precoding vector selection, the interference between users can seriously degrade the overall network data rate. In a close-loop configuration, the base station (BS) receives from each user the preferred precoding vector and modulation and coding scheme (MCS). To achieve the highest multi-user diversity gains and avoid users’ interference, the BS needs to recalculate the precoding vector and MCS for each user. The goal of this paper is to investigate the performance and complexity of state-of-the-art methods for recalculation of precoding vectors such as zero-forcing (ZF) and minimum mean square error (MMSE) for LTE-A scenarios. Additionally, we propose a low-complexity method that jointly recalculates the precoding vector and MCS with a codebook-based approach. We then evaluate our method in terms of throughput and with an LTE simulator.

Keywords— LTE, MU-MIMO, feedback recalculation, ZF, MMSE

1. INTRODUCTION

Multiple input–multiple output (MIMO) techniques are essential features in 3GPP LTE and LTE-A systems in order to achieve high data rates and high system capacity. When a large number of users needs to be served in one cell, high capacity gains can be achieved by transmitting independent data streams to different users sharing the same time-frequency resources. This is called multi-user MIMO (MU-MIMO) and it can be realized through the use of MIMO precoding. Several precoding techniques applicable to the LTE standard have been introduced and discussed in the past few years [1-7].

In a closed-loop configuration, the receiving user obtains downlink channel state information (CSI) by calculating three values that are feedbacked to the base station (BS): channel quality indicator (CQI), rank indicator (RI), and precoding matrix indicator (PMI). With this information, the BS becomes aware of the channel quality of the users and can therefore choose the proper transmit modulation and coding schemes (MCS) for each of them. On the other hand, the PMI shows the precoding vector preferred by the user according to a certain criterion, for example, mutual information. However, these CSI feedback values reported by each user do not consider the interference created to the rest of the users on the same time-frequency resources. The base station should recalculate the PMI and MCS in order to avoid user interference. If we directly apply the CSI values feedbacked by the users in a MU-MIMO transmission, the system performance can be degraded significantly due to interferences between users [7].

With the concerns above, we evaluate in this paper a joint precoding vector and MCS recalculation method which only relies on single user CSI feedback and propose some techniques to reduce its complexity. For this, we first simulate a MU-MIMO transmission from one BS with CSI feedback from different users, and then recalculate MCS with this feedback. Then we adopt Zero Forcing (ZF) or Minimum Mean Square Error (MMSE) algorithms to recalculate the precoding vectors to reduce the interference toward other simultaneously scheduled users. The simulations are based on LTE-A simulator from [8]. Additionally, we discuss about the complexity of the recalculation computation.
The paper is organized as follows. The system model is introduced in section II. The CSI feedback principles are presented in section III, and MCS recalculation methods for normal MU-MIMO, ZF MU-MIMO and MMSE MU-MIMO mode with single user feedback are presented in section IV. Practical considerations and complexity issues are discussed in section V. Simulation results are provided in section VI. Finally section VII draws some conclusions.

II. SYSTEM MODEL

We consider a MIMO-OFDM system with $M$ transmit antennas at BS and $S$ users each one equipped with 1 receive antenna. The input-output relation is given by

$$ y_k = H_k W_k u_k + n_k = G_k u_k + n_k $$

where $u_k$ is the vector of $S$ independent data symbols transmitted in parallel by $M$ transmit antennas and $k$ is the subcarrier index. $W$ is the $MxS$ precoder applied at the BS. The precoding vector feedbacked by the users is codebook-based as defined in LTE standard [9]. $y$ is the vector of signals individually received by the $S$ users and $n$ is the Gaussian noise vector. The $SxM$ matrix $H = [ h_1^T, ..., h_M^T ]^T$ contains the transpose MIMO channel coefficients from the $M$ antennas to the $S$ users.

III. CSI FEEDBACK PRINCIPLES

The general problem of the feedback algorithm consists in choosing the precoder $W_f$ that maximizes the mutual information on each subcarrier $k$, as given by

$$ W_i^* = \text{arg max}_{W_i} \Sigma_{k=1}^K I_k(W_i) $$

where

$$ I_k = \Sigma_{s=1}^S \log_2 \left( 1 + \text{SINR}_k^S(W_i) \right) $$

and $\text{SINR}_k^S$ is the signal-to-interference-plus-noise ratio at subcarrier $k$ for user $s$

$$ S\text{INR}_k^S = \frac{|e_k^S|^2}{\Sigma_{i \neq s} |e_k^i|^2 + \sigma^2} $$

and $W_i$ is the codebook-based precoder vector defined in LTE standard [9] and $\sigma^2$ is the noise power.

For MU-MIMO we consider rank-one feedback, i.e. only one data stream is transmitted to each user. For CQI feedback, as shown in [4] [10], we calculate post-equalization SINR for each subcarrier and map the CQI value to the maximum MCS such that a block error rate (BLER) lower than 0.1 is achieved.

IV. MCS AND PRECODING RECALCULATION AT THE BASE STATION

A. Unitary precoding

In a closed-loop transceiver configuration, each user will feedback CSI to let the BS acquire initial information about users’ channel quality and preferred precoding vector. Then the BS will decide which user to schedule on each time-frequency resource block. In a MU-MIMO transmission, more than one user can be scheduled on the same time-frequency resources, creating interference among users. For example, figure 1 shows 4 possible precoding vectors ($v$) from the precoding codebook. Each user $s$ selects a precoding vector $V_s$ on subcarrier $k$ that is the closest to the Hermitian of the channel vector $h_k^H$ with an angle of departure $\theta_k$. $e_k$ is the error between $h_k^H$ and the precoding vector $v_k^H$. We can express...
\[ h_s = \| h_s \| (v_s + e_s)^H = \| h_s \| \tilde{h}_s. \]

Due to this error, the user selecting \( v_1 \) will create some interference on users selecting \( v_2 \).

For single user rank-one feedback, each user is not aware of other users’ precoding or CQI selection [7], which makes the mapping of a CQI value to a MCS inaccurate. Therefore, CQI recalculation \( CQI^r_{k} \) is necessary at BS. Several recalculation schemes have been discussed in recent years. It is often recalculated as

\[
CQI^r_{k} = \frac{P | h_s v_s |^2}{\sum_{q=1}^{S} | h_q v_q |^2 + 1} = \frac{[1 + e_s v_s]^2 CQI_s}{CQI_k + S [1 + e_s v_s]^2}
\]

where \( P \) is the total power available at the BS and \( S \) is the number of users and \( CQI_s = P | h_s v_s |^2 \) as assumed by the authors in [1]. The numerator represents the signal strength, and the denominator represents the interference from other users. \( V_s \) stands for precoding vector on subcarrier \( k \) selected by user \( s \) and \( V_q \) are precoding vectors for other \( q \) scheduled users on subcarrier \( k \).

Our contribution to reduce complexity in the calculation of equation (5) is to use the feedback CQI index (instead of \( CQI_s = P | h_s v_s |^2 \)) that corresponds to the received SNR and \( \theta_{\text{max}} \) which is known based on the codebook structure. Hence equation (5) can be simply implemented as a lower bound as

\[
CQI^r_{s} \geq \frac{t^2 CQI_s}{\alpha \times CQI_s + St^2}
\]

where \( t = \cos \frac{\theta_{\text{max}}}{2} \) and the factor \( \alpha \) serves to adjust the amount of interference. In this paper we assume it fixed as we only consider a simple 4 vector codebook.

B. Zero forcing precoding

With unitary precoding, the BS transmits data using the precoding vector feedbacked by the user directly without recalculation on the precoder. However, interference due to the precoding vector selection is still present. To improve performance, a pre-equalization step can be applied at transmitter side. For this, the BS calculates a set of beamforming weights in order to maximize the gain towards the user of interest and at the same time minimizing interference towards other simultaneously scheduled users. The principle and the BS operation using ZF is presented in [4]. Let us review the process here.

Step 1: Each user feedbacks its preferred PMI from the precoding codebook. In our case, users choose the PMI that maximizes throughput.

Step 2: Each user reports the CQI such that

\[
CQI^Z_{k} = \frac{P \| h_s \|^2 \cos^2 \theta_s}{1 + P \| h_s \|^2 \sin^2 \theta_s}
\]

which is a simplification of equation (5). We can note that equation (7) needs the computation of the angle \( \theta_s \) between the channel vector and the closest precoding vector, which can only be known if the user feedbacks the channel vector to the BS.
Step 3: We denote with $\hat{H} = [\hat{h}_1^T, ..., \hat{h}_N^T]^T$ the transpose of the quantized channel vector which is closest to a certain precoding vector from the codebook. We then redefine the precoding vector $W$ as

$$W_{ZF} = F \text{ diag}(p)^{1/2} = \hat{H}^H (\hat{H} \hat{H}^H)^{-1} \text{ diag}(p)^{1/2}$$  \hspace{1cm} (8)

Where $p$ is a vector of transmit powers. Thanks to this pre-equalization step at the BS, the CQI does not need to be recalculated. However, to guarantee equal power allocation across the selected users [1][4], the BS estimates the CQI as

$$CQI_s^{ZF,re} = \frac{P_s}{P/S} CQI_s^{ZF} = \frac{P_s}{P/S} \frac{\bar{p}_2 \| f_s \|_2^2 \| h_s \|_2^2 \cos^2 \theta_s}{\frac{\bar{p}_1}{S}} = \frac{S(\| f_s \|_2^2)^{-1} CQI_s \cos^2 \theta_s}{S + CQI_s \sin^2 \theta_s}$$  \hspace{1cm} (9)

where $f_s$ denotes the $s$th column of $F$. In order to reduce the complexity of equation (9), we propose to create a lower bound in function of the CQI index and $\theta_{max}$ such that

$$CQI_s^{ZF,re} \geq \frac{S(\| f_s \|_2^2)^{-1} CQI_s \cos^2 \theta_{max}}{S + CQI_s \sin^2 \theta_{max}}$$  \hspace{1cm} (10)

C. Minimum mean square error precoding

MMSE pre-equalization [2] can also be used to maximize the signal-to-interference plus noise ratio. In this case, the precoding matrix is given as

$$W_{MMSE} = \hat{H}^H (\hat{H} \hat{H}^H + \sigma^2 \times I)^{-1} \text{ diag}(p)^{1/2}$$  \hspace{1cm} (11)

In this case, the BS also has to estimate the CQI for equal power allocation. The calculation is the same as ZF except for the precoding matrix.

V. PRACTICAL CONSIDERATIONS AND COMPLEXITY ISSUES

In a practical scenario, hardware complexity and feedback delay need to be taken into account. For the calculation for CSI feedback, the computational effort can be prohibitively large with a large number of subcarriers and users. To reduce the complexity we suggest the use of channel averaging as shown in [5]. In the following section we use a channel averaging of $72$ subcarriers i.e., $K = 72$. Furthermore, this allows for reduced signaling overhead as less parameters need to be feedbacked.

We also consider the complexity reduction mentioned above, where for the recalculation we use the CQI index, PMI index, and $\theta_{max}$ which is known based on the codebook structure.

VI. SIMULATION RESULTS

In this section we provide link level simulation results to evaluate the performances of the precoding schemes presented in this paper. We also evaluate the performance of CQI and precoding recalculation schemes. For our simulations we use the LTE-A simulator released by Vienna University of Technology [8]. The main simulation parameters are summarized in Table I.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmission bandwidth</td>
<td>1.4MHz</td>
</tr>
<tr>
<td>NFFT</td>
<td>128</td>
</tr>
<tr>
<td>Number of subcarriers</td>
<td>72</td>
</tr>
</tbody>
</table>
### Antenna Configuration

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2x1</td>
<td>(2 users)</td>
</tr>
<tr>
<td>4x1</td>
<td>(4 users)</td>
</tr>
</tbody>
</table>

### Receiver

- Maximal Ratio Combining

### Channel Model

- SUI 1

### Feedback Granularity

- Whole bandwidth

### Channel Estimation

- Ideal

### Codebook

- LTE

### CQI Feedback

- 4 bit

---

### Unitary MU-MIMO

We first consider a 2 user MU-MIMO transmission using the CQI and PMI feedbacked by the user without any recalculation. Figure 2 shows the ideal case that no interference exists between users, which means that the 2 user channel vectors are orthogonal to each other, and so do the precoding vectors. The red line is the 2 user MU-MIMO case and the blue line is our reference single user MIMO (SU-MIMO) case using transmit diversity, i.e. both antennas from the base station transmit the same data to the user. As expected, the BS can transmit at almost double throughput in the MU-MIMO case compared to SU-MIMO when no interference is present. When the user channel vectors are not orthogonal to each other, the selection of a precoding vector from the codebook results in interference among users, with a throughput degradation as shown in figure 3 (low interference case).

In order to reduce the interference among users, CQI recalculation is needed. Here we adopt the recalculation method in section IV.A to improve performance. In figure 4 we can see that MU gain can still be achieved by CQI recalculation in medium and high interference scenarios.
As we increase the interference among users, the throughput degradation becomes severe as seen in figures 4 and 5. By adopting CQI recalculation we can improve performance with low and medium interference, but with high interference we can only achieve a similar throughput as SU-MIMO. Therefore, to achieve spatial diversity gain it is not practical to use unitary precoding. We turn then our attention to ZF and MMSE precoding schemes.

B. ZF MU-MIMO

In this case, the BS recalculates the precoding vector and the CQI as discussed in section IV.B. As we do not need CQI recalculation, we first evaluate the ideal case in which the BS knows exactly the users’ channel vectors and then the realistic case where these channel vectors are mapped to the closest precoding vector from the codebook and feedbacked to the BS. In figure 6 we simulated the ideal case and we can observe that the throughput for the 2-user case is doubled compared to the SU case, and for the 4-user case it is almost four times.

Using the precoding vectors instead of the exact channel vectors, the performance degrades as seen in Figure 7. However, in contrast to the unitary precoding, we can still achieve MU gain. The degradation is evidently larger for 4-user case since the interference increases with the number of users.
c. MMSE MU-MIMO

Finally we evaluate the MMSE case, and also simulated in both ideal and realistic cases as shown in figure 7 and 8, respectively. As expected, the throughput performance is similar to ZF case when the noise power is limited. However, by increasing the noise, we can see that MMSE outperforms ZF precoding scheme as shown in figure 10 and 11 for both 2 user and 4 user case. However, for MMSE the statistical noise information should be feedbacked from the user.

VII. CONCLUSIONS

Three different MU-MIMO precoding schemes with SU feedback are presented in this paper. In particular, CQI and precoding recalculation are derived for each case using rank-one feedback. We also discuss the practical complexity issues and suggest channel averaging and CQI recalculation lower bound to reduce complexity. Through our simulations we show that unitary MU-MIMO precoding can achieve multi user gain in low interference scenarios when CQI recalculation is used. In medium and high interference scenarios, multi user gain can still be achieved by using ZF and MMSE precoding schemes that require an extra pre-equalization step before transmission, with MMSE outperforming ZF with high levels of noise power. This is especially evident in intermediate SNR range since ZF scheme does not take noise into account, while MMSE scheme balances noise and interference.
REFERENCES


Transport Block Scheduling in LTE: Advantages in Structural Limitations

Alessandro Chiumento\textsuperscript{1,2}, Sofie Pollin\textsuperscript{1,2}, Claude Desset\textsuperscript{1}, Liesbet Van der Perre\textsuperscript{1,2}, and Rudy Lauwereins\textsuperscript{1,2}

\textsuperscript{1}Interuniversity Micro-Electronics Center (IMEC) vzw Kapeldreef-75, Leuven, B-3001, Belgium,
\textsuperscript{2}Department of Electrical Engineering ESAT - KU Leuven, Kardinaal Mercierlaan 94, Leuven, B-3001, Belgium

\begin{abstract}

The paper considers the multi-user resource allocation for a practical LTE system. The aim is to exploit multi-user diversity in the resource allocation, yet achieve a homogeneous fixed configuration per user, as determined by the LTE standard. To obtain this, we start from a theoretical OFDMA resource allocation and extend it towards LTE in a second phase. By doing so, we exploit multi-user diversity while meeting LTE standard constraints. To the best of our knowledge, this is the first paper analysing the impact of the LTE physical layer constraint on multi-user resource allocation performance. We show that taking into account realistic LTE limitations a gain can be exploited in both throughput assigned to the user, with an improvement of up to 23\% over state-of-the-art solutions, and power consumption reductions in the order of 10\%.
\end{abstract}

1 Introduction

Improving spectral efficiency has been a constant priority of the wireless community. To achieve high spectral efficiency in multi-user communication systems, the design of the resource scheduling mechanism is crucial. In the context of modern 4\textsuperscript{th} generation telecommunication networks, such as the Long Term Evolution (LTE), improved multi-user resource allocation is even more important in the overall system optimisation, thanks to the many degrees of freedom in resource allocation or frequency planning for these new communication systems [1]. While these flexibility degrees of freedom promise optimal adaptation to fading, interference of multi-user diversity, it however needs to be understood how to optimally exploit them, given realistic system dynamics and standard constraints. Indeed, while LTE allows a detailed channel state reporting, per user, and per resource block, it requires a fixed configuration per user over all the resource blocks assigned to that user [3]. As a result, the resource allocation becomes highly non-linear, and allocating additional resources to a given user might actually result in a lower capacity for that user since the additional resource might have a low channel quality and hence requires a more robust transmission configuration for that user over all allocated resources. The goal of this paper is the introduction of a resource allocation strategy for LTE that takes this into account, yet allows exploiting multi-user diversity and adapts to inter-cell interference.
The vast majority of schedulers are designed to fulfill datarate, fairness or delay constraints [5]. Recently, a lot of attention has been drawn to energy-efficient resource allocation in order to reduce the energy expenditure and networks’ carbon footprints [8]. Solutions are needed however that allow tuning the single resource allocation strategy adaptable towards capacity, energy or fairness, as function of user preferences, application constraints or varying battery budgets. The practical resource allocation problem is hence not only non-linear, it is also multi-dimensional.

Some noticeable work has been done by Videv and Haas in [12] and by Butt et al. in [4] on developing schedulers that would trade off bandwidth with energy. They however make the assumption that a lower transmission rate results in a significantly lower power cost, relying on the convexity of the capacity curve [12, 4]. While this might be true for the theoretical capacity curve, practical power modelling results of LTE base stations have shown that the power cost does not decrease significantly with transmission rate. Because of the losses in the power amplifier and also the fixed power cost, there is only a minimal to no benefit in transmitting at lower rate, since that would require more time to transmit the same amount of data. As a result, while the power to transmit at a lower rate might be slightly lower, the net energy to transmit the same amount of data will be higher due to the longer transmission time. As a result, it is typically assumed that the most energy efficient strategy is to transmit at the highest rate, during the best resources and not waste power on low rate resource blocks [6, 11]. The highest performance is then possible when a cell uses few resources to the most of their capabilities. Scheduling extra resources, with a possible lower channel quality, will hence negatively impact the energy per bit performance of the communication. Clearly, assigning more resources is not always beneficial and this should be considered in the practical resource allocation strategy.

The scope of this work is to show that such a theoretical OFDMA formulation of the scheduling problem, aiming to schedule as much resources as possible, can result in suboptimal capacity and power allocation. An algorithm will be proposed, introducing an extra iteration in the resource allocation, that allows to deliver higher performance or lower power cost by reducing the number of allocated resources when that is relevant. This extra iteration allows the resource allocation mechanism to be aware of the actual performance of the user for that transmission interval on top of the ideal performance determined, per sub-carrier, by any state-of-the-art scheduling algorithm. This awareness moreover allows to trade-off power and throughput performance in the LTE network.

Section II introduces briefly the system and the power model; in section III the problem statement is addressed and in section IV a simple scheduler is proposed. In section V the simulation parameters and results are given, finally, section VI contains the conclusions.

2 System and Power Models

A multi-cell LTE Downlink OFDMA scenario is considered. The system has been simulated using the open source VIENNA system level simulator [9]. The available bandwidth is divided into orthogonal physical resource blocks (PRBs) which represent the smallest granularity the resource allocation manager can assign to each user. The PRB is analog to subcarrier notation in a generic OFDMA formulation. The system bandwidth is flexible and can vary between 1.4 MHz (6 PRBs) and 20 MHz (100 PRBs). For
each time slot, each mobile station (MS) transmits via uplink a channel state indicator (CSI) packet. Each CSI contains a channel quality indicator (CQI) value for each PRB seen by the user. These CQIs are directly linked to the SINR experienced by the user [2]. After PRBs are allocated to the users, the base station, using the adaptive modulation and coding (AMC) strategy decides which modulation and coding scheme (MCS) to use for the transmission that can guarantee a Block Error Rate (BLER) lower than 10%. Different MCSs lead directly to different transmission rates. Table 1 presents the relation between different SINR and CQI values and their respective modulation and coding rate.

In order to quantify the power consumed by each base station a realistic, flexible power model is considered [6]. This model is capable of adapting to various LTE scenarios and accurately scales power figures to match different kind of base stations, e.g., macro, micro, pico and femto, and their conditions of operation eg: MIMO, bandwidth, MCS. The model computes the total power consumed by the base stations by considering the power consumption of each of its main components: the analog RF transceiver, the base band, the power amplifier and the AC-DC/DC-DC conversion and cooling (overhead). The total power consumption is then the sum across the components. Figure 1 shows the typical effect on power consumption of a macro base stations by its components.

Table 1 also shows the relation between total power consumption and MCS adopted by a single sector of a macro base station when the network operates in full load.

It is very important to notice that the base station does not, in fact, assign a different MCS per each PRB but it computes a unique equivalent MCS for the aggregate PRBs assigned to each user, these PRBs form a Transport Block (TB) [3]. This effective transport block MCS is obtained by approximating the TB channel with an equivalent AWGN channel that would allow a BLER lower than 10%. In 2 (a) the

<table>
<thead>
<tr>
<th>SINR</th>
<th>CQI</th>
<th>modulation</th>
<th>code rate (x 1024)</th>
<th>efficiency</th>
<th>Power(W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6.9360</td>
<td>1</td>
<td>QPSK</td>
<td>78</td>
<td>0.1523</td>
<td>292.13</td>
</tr>
<tr>
<td>-5.1470</td>
<td>2</td>
<td>QPSK</td>
<td>120</td>
<td>0.2344</td>
<td>292.55</td>
</tr>
<tr>
<td>-3.1800</td>
<td>3</td>
<td>QPSK</td>
<td>193</td>
<td>0.3770</td>
<td>293.03</td>
</tr>
<tr>
<td>-1.2530</td>
<td>4</td>
<td>QPSK</td>
<td>308</td>
<td>0.6016</td>
<td>293.62</td>
</tr>
<tr>
<td>0.7610</td>
<td>5</td>
<td>QPSK</td>
<td>449</td>
<td>0.8770</td>
<td>294.27</td>
</tr>
<tr>
<td>2.6990</td>
<td>6</td>
<td>QPSK</td>
<td>602</td>
<td>1.1758</td>
<td>294.95</td>
</tr>
<tr>
<td>4.6940</td>
<td>7</td>
<td>16QAM</td>
<td>378</td>
<td>1.4766</td>
<td>295.62</td>
</tr>
<tr>
<td>6.5250</td>
<td>8</td>
<td>16QAM</td>
<td>490</td>
<td>1.9141</td>
<td>296.46</td>
</tr>
<tr>
<td>8.5730</td>
<td>9</td>
<td>16QAM</td>
<td>616</td>
<td>2.4063</td>
<td>297.37</td>
</tr>
<tr>
<td>10.3660</td>
<td>10</td>
<td>64QAM</td>
<td>466</td>
<td>2.7305</td>
<td>298.08</td>
</tr>
<tr>
<td>12.2890</td>
<td>11</td>
<td>64QAM</td>
<td>567</td>
<td>3.3223</td>
<td>299.12</td>
</tr>
<tr>
<td>14.1730</td>
<td>12</td>
<td>64QAM</td>
<td>666</td>
<td>3.9023</td>
<td>300.13</td>
</tr>
<tr>
<td>15.8880</td>
<td>13</td>
<td>64QAM</td>
<td>772</td>
<td>4.5234</td>
<td>301.20</td>
</tr>
<tr>
<td>17.8140</td>
<td>14</td>
<td>64QAM</td>
<td>873</td>
<td>5.1152</td>
<td>302.23</td>
</tr>
<tr>
<td>19.8290</td>
<td>15</td>
<td>64QAM</td>
<td>948</td>
<td>5.5547</td>
<td>303.11</td>
</tr>
</tbody>
</table>

Table 1: SINR and CQI mapping to modulation and coding rate with power consumption at full load
PRBs are assigned to a specific user based on the CQI feedbacks. The line spacing within each assigned PRB is representative of the CQI value associated with that PRB. These PRBs are grouped, in figure 2 (b), forming a transport block. The base station controller checks which modulation and coding rate can be applied to the TB so that the transmission remains reliable. This MCS is then applied identically on all the PRBs composing the TB, figure 2 (c).

The choice of which MCS to use is generally obtained by performing an Effective Exponential Signal-to-noise-ratio Mapping (EESM) \[7, 3\]. The effective transport block SINR \(\gamma_{\text{eff}}\), obtained with EESM, is computed with:

\[
\gamma_{\text{eff}} = -\lambda \log\left(\frac{1}{\text{length}_{TB}} \sum_{k^* \in TB} \exp\left(-\frac{\gamma_{k^*}}{\lambda}\right)\right),
\]

(1)

where \(\lambda\) is a parameter empirically calibrated by the base station as a function of the MCS and \(k^*\) represents the selection of PRBs composing the TB. The throughput of each user, and the power spent by the base station, per transmission interval, then, are not the aggregate ones of the combined PRBs, but they are non-linear functions of the SINRs of the assigned PRBs. The lowest quality CQIs will then dominate the overall TB BLER significantly, and will drive the base station decision towards a lower AMC, reducing so the user’s overall throughput. By adjusting the size of the transport block, removing or adding PRBs, it is then possible to take advantage of the non-linear mapping described above and design a scheduler that can maximise the cell’s rate, minimise its power or be tuned to fit the network’s loads.

3 Problem Statement

It is clear now that performing the scheduling without taking into consideration the actual TB allocation and resulting performance might lead to a misrepresentation of the network capabilities. In this section the differences between a traditional, per PRB only, and the TB aware resource allocation models are presented. Furthermore a framework to explore the possible trade-offs, in terms of achievable datarate versus power consumed in case of TB aware scheduling is introduced.

The generic rate maximisation problem for an LTE downlink cell can be expressed
as:
\[
\max \sum_{x=1}^{X_B} \sum_{k=1}^{K} r_{x,k}(\gamma_{x,k}) \cdot a_{x,k},
\]
(2)

Where $X_B$ is the number of served users and $K$ is the number of available PRBs. $r_{x,k}(\gamma_{x,k})$ is the instantaneous datarate of user $x$ on PRB $k$ determined by the SINR on that PRB and $a_{x,k} = \{1, 0\}$ is a binary variable that indicates whether PRB $k$ has been assigned to user $x$ or not. The problem is then subject to the following constraints:
\[
\sum_{x=1}^{X_s} a_{x,k} = 1, \forall x, \quad (3)
\]
and
\[
\sum_{k=1}^{K} p_k \leq P_{MAX}, \quad (4)
\]

Constraint (3) makes sure that only one user might be allocated on each PRB and constraint (4) makes sure that the sum of the power spent on each PRB is not higher than the base station’s maximum power. For a PRB-only scheduler, the resource allocation is terminated once a solution to (2) has been found.

For a realistic LTE downlink allocation an extra step is necessary. Once the scheduler has allocated the PRB so to fulfil (2), the base station assigns a single MCS per TB and computes $R_{x,TB_x}(\gamma_{eff})$, which is the equivalent rate obtained by user $x$ on its transport block $TB_x$ using formula (1) for the PRBs assigned to the user. This rate is function of the quality of each PRB composing the TB assigned to the user. The optimisation problem then becomes:
\[
\max \sum_{x=1}^{X_B} R_{x,TB_x}(\gamma_{eff}).
\]
(5)

The rate maximisation process is then here redefined in a two-step fashion. Firstly each user is assigned based on the feedback received at PRB level. Secondly, each user’s TB is analysed and only the PRBs which maximise formula (5) are taken into account. The PRBs that didn’t contribute positively to the rate, can either be discarded, reducing the base station’s load or be reassigned to the users in order to further enhance rate performances or fulfil additional scheduling requirements such as a specific quality of service or other scheduling requirements.

4 TB aware scheduling

In this section a simple TB aware implementation of a generic LTE downlink scheduler is presented. The TB choice algorithm functions regardless of the initial PRB assignment, it is, in fact, an overlaid solution and can be implemented on top of any state-of-the-art scheduler.

Algorithm 1 presents the behaviour of this scheduler:
Firstly the users are assigned with the per PRB scheduler.
Phase two: for each scheduled user $x$, determine which PRBs were assigned from the previous phase and their CQIs, here called $PRBS_x$ and $CQIs_x$; these are then reordered from best to worst. In the following loop, at each iteration, one PRB is removed form the available ones starting from the one which carries the lowest CQI (the last) and the TB equivalent CQI, $CQI_{x,TB_x}$, is determined using the EESM formulation. This CQI value is then mapped into the iterative TB Rate $R^i_{x,TB_x}$, the process is non-linear and, thus, it has to be computed at each iteration $i$. Finally, the highest value of $R^i_{x,TB_x}$ is chosen as the TB rate $R_{x,TB_x}$ and all the PRBs that contribute in lowering the overall TB rate are removed from the user’s assigned pool.
**Phase one: Per PRB assignment**

**Phase Two: TB aware assignment**

% loop for each scheduled user

\[
\text{for } x \in X_S \text{ do}
\]

\[
PRB_{x} = \text{PRBs assigned to user } x \text{ from phase one; } CQI_{x} = \text{CQI values relative to } PRB_{x};
\]

reorder \( CQI_{x} \) and \( PRB_{x} \) from best to worst

\[
\text{for } i \in PRB_{x} \text{ do}
\]

% determine TB Rate and Power for varying TB size

\[
CQI_{x,TB} = EESM(CQI_{x}(1 : \text{end} + 1 - i));
\]

\[
CQI_{x,TB} \rightarrow R_{x,TB}^i;
\]

\[
\text{end}
\]

% find max value

\[
\text{find } i \text{ such that } R^i = \max(R);
\]

\[
PRB_{x}(\text{end} + 2 - i : \text{end}) = \text{empty};
\]

\[
\text{end}
\]

**Algorithm 1:** TB aware scheduler

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Macrocells</td>
<td>37</td>
</tr>
<tr>
<td>Sectors per Macrocell</td>
<td>3</td>
</tr>
<tr>
<td>Inter-cell distance</td>
<td>500 m</td>
</tr>
<tr>
<td>Macro antenna gain</td>
<td>15 dB</td>
</tr>
<tr>
<td>Macro Transmit Power</td>
<td>46 dBm</td>
</tr>
<tr>
<td>Macro users per sector</td>
<td>15</td>
</tr>
<tr>
<td>Frequency</td>
<td>2.1 GHz</td>
</tr>
<tr>
<td>System Bandwidth</td>
<td>20 MHz</td>
</tr>
<tr>
<td>Number of PRBs</td>
<td>100</td>
</tr>
<tr>
<td>Access technology</td>
<td>OFDMA FDD</td>
</tr>
<tr>
<td>Number of antennae</td>
<td>1(Tx and Rx)</td>
</tr>
<tr>
<td>Channel model</td>
<td>Winner Channel Model II [10]</td>
</tr>
<tr>
<td>Block fading mean</td>
<td>0 dB</td>
</tr>
<tr>
<td>Block fading deviation</td>
<td>10 dB</td>
</tr>
<tr>
<td>Fast fading</td>
<td>10 dB</td>
</tr>
<tr>
<td>Thermal noise density</td>
<td>-174 dBm/Hz</td>
</tr>
<tr>
<td>Users speed</td>
<td>1 m/s</td>
</tr>
</tbody>
</table>

Table 2: System Parameters

5 Simulation Parameters & Results

Table 2 presents the simulation parameters used to obtain the results presented in this section. Firstly a comparison between the RB only scheduler and the TB aware scheduler is presented. Figures 3 and 4 show the CDF of the datarate achieved by the users and a time plot of the average base station’s power consumption respectively. It can be seen that the TB aware scheduler achieves a small rate improvement but a substantial reduction in power consumption. This is because it anticipates the per user fixed MCS selection by the base station, and optimises the resource allocation with this in mind. This shows that the TB aware scheduler is inherently more efficient,
because it takes into account the relevant constraints. More importantly, the resource allocation could allow tuning towards rate or power by reassigning the PRBs freed in the TB assignment phase to the high throughput users.

Figure 3: CDF of average user throughput for state-of-the-art vs TB-aware scheduler

Figure 4: Average Sector Power Consumption for state-of-the-art vs TB-aware scheduler

To further prove the generality of a TB aware approach we study some of the most well known resource allocation mechanisms used in LTE, and the results are presented in Fig. 5. Six state of the are schedulers have been implemented, namely the round robin, best CQI, proportional fair, Max-Min, resource fair and the iterative Hungarian schedulers, as presented in [5]. Figure 5 shows the average sector’s throughput over the average sector’s power gain for the different resource allocation mechanisms. In order to guarantee a fair representation, the power results obtained by each scheduler are expressed as fractions with respect to the maximum power consumed by a base station in full load with highest MCS. The empty markers represent the schedulers in their PRB only configuration and the full markers the TB aware version of the same schedulers. Every scheduler presents an improvement when TB aware allocation is considered, showing that taking into consideration the intrinsic limitations of the network can play a great role in determining the final user’s performance. The table contained in figure 6 shows the improvements in throughput and power for each scheduler.

Figure 5: Throughput over Power comparison for different state-of-the-art schedulers with and without TB awareness

Figure 6: Improvements in datarate and power consumption for the different implemented schedulers

The round robin scheduler shows the largest gains. This comes from the fact that it randomly assigns PRBs to users, leading to more diversity in CQI values for the
PRBs of a given user. Hence, suppressing low-CQI PRBs enables a significant gain of the TB effective CQI while not wasting power on poor PRBs. The improvement seen over its performances caused by the TB aware scheduling is directly dependent on the amount of poorly assigned PRBs present in the first phase of the resource allocation. All the other schedulers present decisive improvement as well. The Iterative Hungarian scheduler assigns PRBs to users such that each user receives a mixture of high quality and low quality resources. The improvement in datarate given by the TB awareness is indicative then of how influential the EESM mechanism is on the user’s throughput. The TB aware scheduler will always improve over a state-of-the-art one, the improvement would be avoided only if each allocated user would witness perfect channel quality on each PRB. The TB aware addition allows then the resource allocation to be carried more efficiently, achieving superior datarate while using less resources.

6 Conclusions

Traditional resource allocation mechanisms, designed for generic OFDMA networks, do not take into consideration the limitations imposed by the 3GPP LTE standard and thus cannot take full advantage of the network capabilities within that standard. In this work the transport block scheduling problem has been defined and a generic solution has been presented. It has been shown that there is considerable gain, in both datarate and power, that can be exploited over any of the state-of-the-art schedulers analysed.

Bibliography

[8] EARTH. EARTH project deliverable 2.1, "Economical and ecological impact of ICT".
An OMP algorithm with memory for spectrum sensing using the Modulated Wideband Converter

Jean-François Determe†  André Bourdoux†  François Horlin†
† Brussels School of Engineering (Université Libre de Bruxelles) — Avenue F. D. Roosevelt 50, Brussels, B-1050, Belgium
‡ Interuniversity Micro-Electronics Center (IMEC) vzw — Kapeldreef 75, Leuven, B-3001, Belgium
Email: {jdeterme, fhorlin}@ulb.ac.be, bourdoux@imec.be

Abstract

The ever-improving Orthogonal Matching Pursuit (OMP) algorithms have enabled one to deal with a wide variety of \( \ell_0 \)-norm problems more and more efficiently. Among all of the applications of OMP, the identification of active frequency bandwidths within sparse spectrums has remained a challenge due to the reliability to be achieved. This particular application is generally associated with the so-called cognitive radios which aim at detecting available bands within sparse multi-gigahertz wireless spectrums to transmit their own data streams. This paper presents a novel OMP algorithm called EDA-OMPMMV exhibiting memory capabilities which prevent bad detections from occurring by averaging the results of successive iterations. This algorithm has been especially designed for the Modulated Wideband Converter (MWC) and exhibits optimal performance for spectrums undergoing limited modifications over the time.

Index Terms

Orthogonal Matching Pursuit, Modulated Wideband Converter, Multiple Measurement Vector, OMPMMV, EDA-OMPMMV.

I. INTRODUCTION

The increasing need for communication bandwidth has led to a saturation of the available spectrum since almost every frequency band is now allocated to specific applications. This phenomenon complexifies the development of new communication systems (e.g. 4G mobile communication systems). Whilst it could be possible to "recycle" allocated frequency bands to use them in conjunction with more performant technologies, it is generally not a viable option since all the devices operating on the existing system cannot handle new communication technologies (e.g. 2G/3G mobile phones cannot operate on 4G networks). Moreover, using higher carrier frequencies is not always possible for economical or power-consumption reasons.

However, it is worth pointing out that, at a given time instant and location, only a small percentage of the allocated frequency bands is used [2]. It is therefore of interest to consider using these free frequencies to transmit data by means of the so-called cognitive radios.

Cognitive radios are devices capable of scanning wideband radio spectrums to detect the available frequency bands in order to superimpose new communications signals. This task requires the use of advanced signal detectors. The aim of the present paper is to design a new variety of algorithms that can be used in conjunction with the Modulated Wideband Converter [3] so as to achieve spectrum sensing at sub-Nyquist rates.

The novel algorithm is denoted as Exponential Decay Averaging Orthogonal Matching Pursuit Multiple Measurement Vector (EDA-OMPMMV). It averages the results of previous iterations so as to increase
II. SYSTEM MODEL

The Modulated Wideband Converter (MWC) is a device capable of sampling frequency-sparse signals at sub-Nyquist rates. As depicted in Figure 1, the MWC consists of \( m \) channels multiplying the incoming wideband signal \( x(t) \) with \( T_p \)-periodic functions \( p_i(t) (f_p = 1/T_p) \), low-pass filtering the resulting signals and sampling them at a low rate \( f_s \) such that no aliasing occurs (i.e. the low-pass filter admits a two-sided bandwidth equal to \( f_s \)).

\( p_i(t) (i = 1, \ldots, m) \) are binary piece-wise constant functions so that they are easily implemented using CMOS technologies. Generally, special codes like Gold codes of appropriate order are injected within each period of these functions. Therefore, the commutation (or chipping) frequency of functions \( p_i(t) \) is equal to \( f_p M \) where \( M \) is defined as the length of each code sequence.

The operations performed by the MWC can be mathematically expressed by Equations (1)-(3) [3]:

\[
y(f) = A \cdot \tilde{z}(f)
\]  

(1)

where

\[
\tilde{z}(f) = [X(f - L_0f_p) \ldots X(f + L_0f_p)]^T
\]  

(2)

\[
y(f) = [Y_1(e^{j2\pi fT_s}) \ldots Y_m(e^{j2\pi fT_s})]^T
\]  

(3)

\( \tilde{z}(f) \) can be seen as the decomposition of the incoming wideband signal spectrum \( X(f) \) into frequency bands of bandwidth \( f_s \) distant of \( f_p \) hertz from each other. As for \( y(f) \), it gathers the Discrete-Time Fourier Transforms (DTFT) of each sampled channel.

Matrix \( A \) depends on the MWC configuration. Parameters of interest are:
1) The frequency of repetition \( f_p \) of functions \( p_i(t) \)
2) The codes used for each \( T_p \)-periodic function \( p_i(t) \) (i.e. type and length of the code)
3) The sampling frequency $f_s$

The problem to be solved is to estimate the frequency support of the incoming wideband signal $x(t)$ assumed to be sparse in the frequency domain. This is equivalent to finding a sparse support for $\mathbf{z}$ while ensuring that Equation 1 holds true.

For the single purpose of determining active frequency bands, Problem (1) can be recasted into Problem 5. Let $Q$ be defined as

$$Q = \sum_{n=-\infty}^{+\infty} y[n] \cdot y[n]^H$$

Equation (4)

Let Equation 5 define a Multiple Measurement Vector (MMV) problem:

$$Q = A \cdot U$$

Equation (5)

Please note that the infinite sum in Equation (4) can actually be limited to a finite number of terms in practice. The number of terms in the sum should be sufficiently high to ensure that the columns of $Q$ form a frame for all the possible realizations of $y[n]$. This number of terms will increase as the spectrum becomes less sparse since it means that possible values of $y[n]$ are more diverse and require to use more vectors when computing $Q$ to be completely mapped.

It can be proven that the frequency support of the incoming wideband signal $x(t)$ (which is assumed to be sparse) can be obtained by solving Problem (6):

$$U^* = \arg\min_U (\|U\|_0) \quad \text{subject to} \quad Q = A \cdot U$$

Equation (6)

where $U$ is the solution obtained for Problem (6), one can prove [3] that the support of $U$ defined as $\text{supp}(U) = \bigcup_i \text{supp}(U_i)$ is the same as that of the wideband spectrum provided that it is sparse enough. In other words, $\text{supp}(\mathbf{z}) = \text{supp}(U)$.

Assuming $U$ denotes the solution obtained for Problem (6), one can prove [3] that the support of $U$ defined as $\text{supp}(U) = \bigcup_i \text{supp}(U_i)$ is the same as that of the wideband spectrum provided that it is sparse enough. In other words, $\text{supp}(\mathbf{z}) = \text{supp}(U)$.

Nevertheless, $\ell_0$-norm problems are known to be NP-Hard [1] which means that appropriate heuristics should be used. The Orthogonal Matching Pursuit Multiple Measurement Vector (OMPMMV) algorithm is a computationally tractable heuristic dealing with such problems. Classically, OMPMMV is used to compute the spectrum occupancy over the time without taking into account results obtained for past iterations. Our contribution consists in extending the classical OMPMMV algorithm so that the estimate of the spectrum occupancy at one time instant also depends on past estimates. This helps to improve the performance of the frequency support recovery when the spectrum is invariant. The cost to be paid is transients occurring when the spectrum is modified.
III. (EDA-)OMPMMV ALGORITHM

Let $\mathbf{A} = [\varphi_1 \ldots \varphi_L]$ where $\varphi_i$ are column vectors.

The classical OMPMMV algorithm is described hereafter [1].

**Classical OMPMMV algorithm:**

1) Initialization: residual $R_0 = \mathbf{Q}$ and support $S_0 = \emptyset$

2) At step $t$ (repeat until criterion is met)
   a) Choose $\varphi_{k_t} = \underset{\varphi_k}{\text{argmax}} \left( \| R_{t-1}^H \varphi_k \|_2 \right)$
   b) $S_t = S_{t-1} \cup \{ k_t \}$; $Y_t = \mathbf{A}_{S_t} \mathbf{A}_{S_t}^+ R_{t-1}$
   c) $R_t = R_{t-1} - Y_t$

**Output:** Support $S$ at last step

The algorithm first computes the correlation of the residual with each of the normalized atoms $\varphi_k$ in matrix $\mathbf{A}$. It is then assumed that the atom $\varphi_{k_t}$ exhibiting the highest scalar product with the residual should be chosen. The support is then updated and the projection of the residual $R_{t-1}$ onto the direction spanned by the chosen atom is computed by means of the orthogonal projector $\mathbf{A}_{S_t} \mathbf{A}_{S_t}^+$. The residual is finally updated by subtracting the previous projection in such a way that the new residual is orthogonal to the subspace spanned by atom $\varphi_{k_t}$. The process is repeated until a given number of iterations is reached or until the $\ell_2$-norm of the residual is lower than a prescribed threshold.

Please note that choosing the atom exhibiting the highest correlation is not the best choice in the general case. However, it can be proven that it is statistically a good choice provided that matrix $\mathbf{A}$ satisfies the so-called RIP (Restricted Isometry Property) criterion with a small restricted isometry constant $\delta$ [5]. This criterion is satisfied with high probability when using appropriate codes (e.g. Gold codes) for functions $p_i(t)$ [4].

The Restricted Isometry Property (RIP) describes the ability of a matrix $\mathbf{A}$ to conserve the $\ell_2$-norm. It can be formulated as in Equation (9) [1].

\[
(1 - \delta) \| y \|_2^2 \leq \| \mathbf{A} \cdot y \|_2^2 \leq (1 + \delta) \| y \|_2^2
\]

where $y$ is a vector of appropriate dimensions. $\delta$ is called the restricted isometry constant and quantifies the invariance of norm $\ell_2$ attained by matrix $\mathbf{A}$. It is important to notice that low values of $\delta$ increase the reliability of the heuristics dealing with the minimization of norm $\ell_0$.

As already pointed out, the flaw of OMPMMV is that past results are never used although the spectrum should not undergo high variations for two consecutive iterations. Our contribution consists in building a novel version of the classical OMPMMV algorithm that accounts for previous iterations so as to mitigate bad choices of the chosen atoms in OMPMMV. The proposed algorithm is referred to as Recursive Exponential Decay Averaging OMPMMV (EDA-OMPMMV).

The proposed heuristic consists of two parts:

1) The classical OMPMMV algorithm is used but returns the correlations $\| R_{t-1}^H \varphi_k \|_2$ for each atom $\varphi_k$ instead of the final support $S$.

2) The newly computed correlations are averaged with previous iterations such that former contributions vanish quickly. A choice is then made on the basis of the averaging result.
The complete algorithm is detailed hereafter:

- **Initialization**: averagedCorr = vector of zeros;
- **Repeat at regular intervals**:
  1) **Classical OMPMMV algorithm**:
     a) Initialization: residual $R_0 = Q$ and support $S_0 = \emptyset$
     b) At step $t$ (repeat until criterion is met)
        i) Choose $a_k = \arg\max_{a_k} \left( \| R_{t-1}^{H} a_k \|_2 \right)$
        ii) $S_t = S_{t-1} \cup \{k_t\}$; $Y_t = A_{S_t} A^{+} R_{t-1}$
        iii) $R_t = R_{t-1} - Y_t$
     Output: newCorrs: Correlations for each atom (correlation of unused atoms is set to 0)
  2) **Averaging procedure**:
     a) Update averagedCorr: averagedCorr = newCorrs + $g$*averagedCorr where $g$ is called the forgetting factor and belongs to $[0; 1[ \subset \mathbb{R}$. Please note that $g$ is fixed beforehand and its choice depends on several parameters described hereafter.
     b) Choose $m$ frequency bands that correspond to the highest entries in vector averagedCorr.

Please remark that the proposed estimator remains bounded as long as $g \in [0; 1[ \subset \mathbb{R}$.

**Proof**: Clearly, each component of newCorrs remains bounded by a positive real number $\delta$. Hence,

$$\text{averagedCorr}|_i \leq \sum_{i=0}^{n} \delta g^i = \delta \frac{1 - g^n}{1 - g} \leq \delta \frac{1}{1 - g} < g \neq 1 \infty$$

where $n$ is the current iteration and averagedCorr|$i$ denotes the $i$-th component of vector averagedCorr. The first equality corresponds to the sum of a geometric series while the inequality $\leq$ holds true because $0 \leq g \leq 1$.

The choice of adding memory to the algorithm provides better filtering of the abnormalities but also comes along with transients when the spectrum availability is modified. Since EDA-OMPMMV uses previous values of the correlations, time is needed to make old contributions to vanish (they vanish more quickly as $g$ is set low). Hence, the proposed algorithm introduces a **trade-off between robustness in stationary situations** (filtering of abnormalities) and **convergence speed when the spectrum is modified**. This compromise can however be efficiently parametrized by means of parameter $g$.

The optimal value of $g$ depends on several properties:
- The power and nature of the noise applied to the incoming wideband signal $x(t)$. The more powerful the noise, the higher $g$ should be in order to average the noise over longer time duration.
- The properties of matrix $A$. Let matrix $A$ admit a restricted isometry constant $\delta$ that is low. In such a case, $g$ can be lowered since choosing the highest correlations $\| R_{t-1}^{H} a_k \|_2$ in the heuristic becomes a choice that is statistically more efficient.
- The number of channels $m$. The number of bands that can be activated is actually equal to $m$ so that a high number of channels means that it is less likely that a band of interest is missed by the heuristic. Moreover, $m$ being a dimension of matrix $A$, the restricted isometry constant $\delta$ of $A$ can be positively or negatively affected by the choice of $m$.
- The average frequency at which the spectrum is subject to modifications. If the environment is prone to frequent alterations, it is then of interest to minimize $g$ as high values of the forgetting factor come along with long transients.
In practice, the value of $g$ can be finetuned by means of numerical simulations or experiments in the real environment.

Please note that EDA-OMPMMV boils down to OMPMMV if $g = 0$.

IV. SIMULATION RESULTS

This Section aims at evaluating how using EDA-OMPMMV instead of OMPMMV improves the results obtained for the frequency support detection. Each individual simulation case consisted in sending symbols of duration $T_{\text{symb}} = 1/(10\text{MHz}) = 100\text{ns}$ on different carrier frequencies. The simulation time for each case was equivalent to $200 * T_{\text{symb}} = 20\mu\text{s}$.

The spectrum is estimated every $T_p = 100$ nanoseconds by means of OMPMMV and EDA-OMPMMV. The estimations are conducted as soon as the first symbol has been sent. Hence, one may expect some transients occurring at the beginning of the simulation.

The spectrum undergoes variations at times $50 * T_{\text{symb}}$, $100 * T_{\text{symb}}$, $150 * T_{\text{symb}}$. These variations consist in the replacement of one stream of symbols by another one that is carried at a different frequency. The number of channels $m$ has been chosen equal to 9.

Figure 2 depicts the performance achieved by OMPMMV and EDA-OMPMMV for a single case when determining the frequency support of the wideband signal. Every error corresponds to one component of $\hat{z}$ that has not been activated although it should have been in an ideal case. The number of errors is computed at each symbol time $T_p$.

As expected, transients occur for EDA-OMPMMV when the spectrum undergoes variations (i.e. at time instants $50 * T_p$, $100 * T_p$, and $150 * T_p$) and also when the algorithm is initialized. Although these transients decrease the performance achieved by EDA-OMPMMV, their influence is counterbalanced by the stability of the frequency support recovery for constant spectrums.

The first simulation scenario has been conducted by generating 50 individual cases for each 2-tuple $(g, \text{SNR})$. The complete simulation time for a single case is set equal to 200 symbol duration times 200.
(i.e. 200 * \( T_p \)). The spectrum estimation is performed each \( T_p \) seconds. As already stated previously, the spectrum undergoes 3 alterations occurring at times 50 * \( T_p \), 100 * \( T_p \) and 150 * \( T_p \).

Figure 3: OMPMMV versus EDA-OMPMMV – 1st case (3 spectrum alterations) – \( g \in \{0.6, 0.65, \ldots, 0.95\} \) — \( m = 9 \) — SNR \( \in \{0, 1, 2, \ldots, 12\} \) dB

Figure 3 depicts the performance of both OMPMMV and EDA-OMPMMV when determining the frequency support of the incoming wideband signal. Please note that any probability that is lower than 75 % has been represented by black areas. This choice provides better contrast.

Also remark that OMPMMV should not be affected by the forgetting factor \( g \). However, due to the limited number of cases that are generated, some variations in the probability of full frequency support recovery occur.

Although the number of simulated cases is limited, Figure 3 clearly demonstrates that using EDA-OMPMMV instead of OMPMMV for low-SNR scenarios drastically improves the frequency support recovery reliability. It is interesting to know that the OMPMMV probability of full support recovery is actually very low (less than 50 %) for SNRs equal to 0 dB.

The second simulation case is very similar to the first one. The only difference is the number of spectrum alterations that occur. The frequency support is indeed only modified once at time instant 100 * \( T_p \).

Figure 4: OMPMMV versus EDA-OMPMMV – 2nd case (1 spectrum alteration) – \( g \in \{0.6, 0.65, \ldots, 0.95\} \) — \( m = 9 \) — SNR \( \in \{0, 1, 2, \ldots, 12\} \) dB
The results obtained in Figure 4 are similar to those of Figure 3. EDA-OMPMMV performance has however increased since the number of frequency support alterations has been reduced.

Figure 5: EDA-OMPMMV full support recovery probability – Case 1 versus Case 2 – \( g \in \{0.6, 0.65, \ldots, 0.95\} \) — \( m = 9 \) — SNR 0 dB

Finally, Figure 5 depicts the probability of full frequency support detection for a SNR of 0 dB. The main observation is that the optimal value of \( g \) is a function of the frequency at which spectrum alterations are produced. For Case 1, the optimal value of \( g \) is close to 0.85 while it is located near 0.9 for Case 2.

V. CONCLUSION

First of all, the Modulated Wideband Converter (MWC) and its main properties have been described. Then, the regular OMPMMV algorithm to be used in conjunction with the MWC has been introduced. Moreover, a novel OMPMMV algorithm (referred to as EDA-OMPMMV) has been proposed. EDA-OMPMMV includes memory so as to provide better filtering of the noise while having a limited impact on the performance when the spectrum undergoes modifications. The forgetting factor \( g \) of EDA-OMPMMV enables one to finetune the memory of the algorithm according to the rate at which spectrum alterations occur.

Finally, a campaign of simulations has proven that EDA-OMPMMV outperforms OMPMMV for spectrums exhibiting frequency support modification rates that are low enough. In particular, it has been highlighted that low values of the forgetting factor should correspond to less frequent spectrum alterations so as to achieve optimal performance.

REFERENCES

A Low-Complexity Algorithm for Sum Rate Maximization in Multi-Cell Opportunistic DF Relayed OFDMA Downlink Systems

Zhiwen Jin†, Tao Wang‡, Ji-Bo Wei♭, Jérôme Louveaux† and Luc Vandendorpe†

†ICTEAM institute, Université catholique de Louvain, Louvain-la-Neuve 1348, Belgium.
‡School of Communication and Information Engineering, Shanghai Univ., Shanghai 200072, China.
♭School of Electronic Science and Engineering, National Univ. of Defense Tech., Changsha 410073, China.

Abstract

This paper considers a multi-cell OFDMA downlink system aided by several relay stations (RSs). The opportunistic DF protocol proposed by [1] is applied. The system sum rate maximization problem is formulated with a total power constraint in each cell. Note that, when modeling the inter-cell CCI of a subcarrier in a selected cell, instead of using an additional integer variable to indicate whether a station in an interfering cell transmits data on this subcarrier or not, we use the corresponding power value to do it. This choice is motivated to simplify the system sum rate expression and facilitate the algorithm design. An iterative low-complexity resource allocation (RA) algorithm is proposed to optimize mode selection (decision whether relaying should be used or not and which relay), subcarrier assignment (MSSA) and power allocation (PA) alternatively. During the MSSA stage of each iteration, the formulated problem is decoupled into subproblems which can easily be solved in linear time. During the PA stage, an algorithm based on single condensation and Lagrange duality (SC-LD) is designed to optimize PA with the tentative MSSA results. Through numerical experiments, the convergence of the low-complexity algorithm (LCA) as well as its benefit compared with a centralized algorithm (CA) are illustrated.

Index Terms

OFDMA, opportunistic DF relaying, resource allocation, cochannel interference mitigation.

I. INTRODUCTION

In next generation wireless communication networks, high data rate and ubiquitous coverage are strongly required. To achieve this goal, relaying technologies are recommended due to its potential feature for coverage extension and data rate improvement [2]. On the other hand, orthogonal frequency division multiple access (OFDMA) receives much attention due to its inherent ability to combat frequency-selective multi-path fading and its flexibility in applying dynamic radio resource allocation (RA) for performance improvement. Thus, it is quite promising to combine relaying technologies with OFDMA in future cellular systems.

Considering relay aided OFDMA systems, [3] and [4] have investigated an always-relaying decode-and-forward (DF) protocol where transmissions are always aided by the relay station (RS). To better exploit the flexibility in system design, we have proposed an opportunistic DF protocol in [1] where the transmission at a subcarrier is carried out either with or without the help of the RS. In this paper, we adopt this protocol in multi-cell DF relayed OFDMA systems with full frequency reuse, where cochannel interference (CCI) becomes a key factor affecting the system performance. In order to optimize the system sum rate, RA based CCI mitigation will be considered here.
Concerning RA problems in multi-cell DF relayed OFDMA systems, [5] and [6] have recently proposed joint RA schemes to maximize the sum rate over all cells and the weighted sum of per cell min-rate respectively. However, the proposed centralized algorithms (CAs) seem to be quite heavy to implement in practice.

Compared with the above existing works, the contributions of this paper are as follows:

- We formulate the sum rate maximized joint RA problem in multi-cell OFDMA downlink systems aided by opportunistic DF RSs in each cell. Note that, when modeling the inter-cell CCI of a subcarrier in a selected cell, instead of using an additional integer variable to indicate whether a station in an interfering cell transmits data on this subcarrier or not, we use the corresponding power value to do it. This choice is motivated to simplify the system sum rate expression and facilitate the algorithm design.

- We propose an iterative low-complexity RA algorithm to optimize the mode selection, subcarrier assignment (MSSA) and the power allocation (PA) alternatively with the sum rate keeping increasing. During the MSSA stage, the optimization problem is decoupled into subproblems which can be solved in linear time. During the PA stage, the algorithm based on single condensation and Lagrange duality (SC-LD) is carried out iteratively. As will be illustrated through numerical results, the proposed algorithm converges quite fast and is more practical compared to the CA of [6].

The rest of this paper is organized as follows. First, the system model and the problem formulation for the considered system are presented in the next section. The proposed algorithms are described in section III. The convergence of the proposed algorithm and its benefit compared with a CA are illustrated by numerical experiments in section IV. Finally, conclusions are drawn in section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

We consider a multi-cell OFDMA system with $N$ cells. Each cell contains $R$ DF RS aiding downlink transmission from a base station (BS) to $U$ mobile stations (MSs) over $K$ parallel subcarriers. Over each subcarrier, a symbol is transmitted in either relay aided mode or direct mode. If the relay aided mode is selected, the symbol is first broadcast by the BS over a subcarrier $k$ during the first time slot (TS). Then a selected RS decodes the received symbol and relays it to the allocated MS over the same subcarrier $k$ with the BS keeping quiet during the second TS. Note only one relay is helping here. The destination only decodes the symbol received during the second TS (meaning there is no combination between the information transmitted from a base station (BS) to mobile stations (MSs) over all subcarriers.

During the first TS, it is then decoded by the targeted MS. The BS keeps silent during the second TS. Note that, instead of using an additional integer variable to indicate whether a station in an interfering cell $n'$, a symbol $\sqrt{P_{s_n}} x_n^k$ is also produced from the interfering BS $s_n'$ at the same subcarrier. This choice is motivated to simplify the system sum rate expression and facilitate the algorithm design. The targeted MS $d_{un}$ receives signals from all BSs during the first TS. After decoding the received signal, the achievable rate for subcarrier $k$ when allocated to $d_{un}$ is given by

$$R_{un,1}^k (P) = \log \left(1 + \Gamma_{d_{un},t_1}^k\right)$$  \hspace{1cm} (1)

in nats/two-TSs, where

$$\Gamma_{d_{un},t_1}^k = \frac{P_{s_n} G_{s_n}^k}{\sigma^2 + \sum_{n',n'' \neq n} P_{s_{n'}} G_{s_{n'}}^k}$$ \hspace{1cm} (2)

denotes the signal-to-interference-plus-noise ratio (SINR) associated with the decoding of $x_n^k$ at $d_{un}$ during the first TS. $\bar{P}_{d_{un},t_1} = \sigma^2 + \sum_{n',n'' \neq n} P_{s_{n'}} G_{s_{n'}}^k$ denotes the sum power of the AWGN and the interferences...
received by \( d_{un} \). \( G_{jn}^k \) denotes the channel gain for subcarrier \( k \) from \( s_n \) to \( d_{un} \).

Let us now consider the relay aided data transmission when subcarrier \( k \) is allocated to MS \( d_{un} \) and helped by a RS \( r_{jn} \). A symbol \( \sqrt{P_{sn}^k x_n^k} \) is first produced by \( s_n \) at subcarrier \( k \) during the first TS, while \( r_{jn} \) keeps quiet. At the same time, in an interfering cell \( n' \), a symbol \( \sqrt{P_{s_{n'}}^k x_{n'}^k} \) is also produced from the interfering BS \( s_{n'} \) at the same subcarrier. At the end of the first TS, the SINR associated with decoding the received symbol at \( r_{jn} \) is expressed by

\[
\Gamma^k_{r_{jn}} = \frac{P_{s_n}^k G_{s_n,r_{jn}}^k}{f^k_{r_{jn}}}
\]

where \( f^k_{r_{jn}} = \sigma^2 + \sum_{n',n'\neq n} P_{s_n}^k G_{s_n,r_{jn}}^k \) denotes the sum power of the AWGN and the interferences received by \( r_{jn} \). \( G_{s_n,r_{jn}}^k \) denotes the channel gain of subcarrier \( k \) from \( s_n \) to \( r_{jn} \).

All BSs keep silent during the second TS, while \( r_{jn} \) reencodes the decoded symbol and forwards \( \sqrt{P_{r_{jn}}^k x_{n}^k} \). Here \( P_{r_{jn}}^k \) denotes the transmit power allocated to \( r_{jn} \) at subcarrier \( k \). At the same time, in the interfering cell \( n' \), \( r_{jn'} \) also forwards \( \sqrt{P_{r_{jn'}}^k x_{n'}^k} \) to retransmit a symbol \( x_{n'}^k \). At the end of the second TS, the SINR associated with decoding the received symbol at \( d_{un} \) is expressed by

\[
\Gamma^k_{d_{un},t_2} = \frac{P_{r_{jn}}^k G_{r_{jn},d_{un}}^k}{f^k_{d_{un},t_2}}
\]

where \( f^k_{d_{un},t_2} = \sigma^2 + \sum_{n',n'\neq n,j'\neq 1} P_{s_{n'}}^k G_{s_{n'},r_{jn'}}^k G_{r_{jn'},d_{un}}^k \) denotes the sum power of the AWGN and the interferences received by \( d_{un} \). \( G_{r_{jn},d_{un}}^k \) denotes the channel gain of subcarrier \( k \) from \( r_{jn} \) to \( d_{un} \).

The maximum achievable rate for subcarrier \( k \) when in relay aided mode and allocated to \( d_{un} \) is then given by [7]

\[
R_{u_{jn},2}^k (P) = \min \left( \log \left( 1 + \Gamma^k_{r_{jn}} \right), \log \left( 1 + \Gamma^k_{d_{un},t_2} \right) \right)
\]

in nats/two-TSs.

**B. Problem Formulation**

In order to formulate the RA problem, two binary variables \( a_{jn}^k \) and \( b_{jn}^k \) are introduced to describe the mode selection and the subcarrier assignment in the two TSs. Specifically, \( a_{jn}^k = 1 \) indicates that subcarrier \( k \) is allocated for data transmission to \( d_{un} \) in direct mode. \( b_{jn}^k = 1 \) indicates that subcarrier \( k \) is allocated for data transmission to \( d_{un} \) aided by \( r_{jn} \).

We consider maximizing the system sum rate under per cell total power constraints. The optimization variables are the transmission mode for each subcarrier, the subcarrier assignments and the subcarrier power allocations at the BSs and the RSs. According to the system model, the considered RA problem can be formulated as:

\[
\max_{P, A, B} R(P, A, B) \\
\text{s.t.} \\
C1 : a_{jn}^k \in \{0, 1\}, b_{jn}^k \in \{0, 1\}, \forall u, j, n, k, \\
C2 : \sum_{u=1}^U a_{jn}^k + \sum_{u=1}^U b_{jn}^k \leq 1, \forall n, k, \\
C3 : P_{s_n}^k \geq 0, P_{r_{jn}}^k \geq 0, \forall j, n, k, \\
C4 : \sum_{k=1}^K P_{s_n}^k + \sum_{k=1}^K \sum_{u=1}^U \sum_{j=1}^J b_{jn}^k P_{r_{jn}}^k \leq P_{T}, \forall n,
\]
where $\mathbf{P} = \{P^k_n, P^k_{r,j,n}\}$, $\mathbf{A} = \{a^k_{u,n}\}$, $\mathbf{B} = \{b^k_{u,j,n}\}$ and
\[
R(\mathbf{P}, \mathbf{A}, \mathbf{B}) = \sum_{n=1}^{N} \left( \sum_{k=1}^{K} \sum_{u=1}^{U} a^k_{u,n} R^k_{u,n,1}(\mathbf{P}) \right) + \sum_{k=1}^{K} \sum_{u=1}^{U} \sum_{j=1}^{J} b^k_{u,j,n} R^k_{u,j,n,2}(\mathbf{P}) .
\]
(6)

Here, $P^n_{m}$ denotes the available sum power in cell $n$. $C1$ and $C2$ ensure that each subcarrier $k$ in each cell $n$ can select only one mode (direct/relay-aided) to transmit data towards only one MS $d_{u,n}$ with the help of only one RS. Moreover, $C3$ and $C4$ ensure that the consumed sum power for each cell is less than the available sum power. This type of power constraints gives an upper bound of the system performance. In practice, each station (BS, RS, MS) in each cell will have an individual power constraint.

III. ALGORITHM DEVELOPMENT

A. Algorithm Overview

In order to solve problem (5), an iterative coordinate ascent (CA) approach is adopted. Each iteration is carried out in two stages: the MSSA stage and the PA stage. We introduce integer $m$ to indicate the iteration number and add superscript $m$ to variables obtained at the end of iteration $m$. Specifically, we first set $m = 0$ and initialize the binary variables by using the optimal algorithm when CCI is ignored.

During the MSSA stage of iteration $m$, problem (5) is decoupled into subproblems with $\hat{\mathbf{P}} = \mathbf{P}^m$, which can easily be solved in linear time. After optimization, $\{\mathbf{A}^m, \mathbf{B}^m\}$ will be obtained.

During the PA stage of iteration $m$, a low-complexity PA algorithm based on single condensation and Lagrange duality (SC-LD) is used to solve problem (5) with $\mathbf{A} = \mathbf{A}^{m-1}$ and $\mathbf{B} = \mathbf{B}^{m-1}$. As will be shown in subsection III-C, the SC-LD algorithm converges to a point satisfying the Karush-Kuhn-Tucker (KKT) conditions of problem (5) with $\mathbf{A} = \mathbf{A}^{m-1}$ and $\mathbf{B} = \mathbf{B}^{m-1}$. After convergence, the power vector $\mathbf{P}^m$ will be output.

Finally, we stop the iterations when the sum rate increase is below a prescribed value $\epsilon_1$. The CA based RA algorithm keeps on pushing up the system sum rate and converges to a local optimum. Note that each stage is carried out with analytical equations. Thus the proposed RA algorithm has a lower-complexity compared to the CA of [6].

B. MSSA Optimization

In this subsection, the MSSA stage for iteration $m$ is considered. After setting $\hat{\mathbf{P}}$ to $\mathbf{P}^m$, it is important to note that each subcarrier in each cell has independent constraints. Therefore problem (7) can be decoupled into $N \times K$ subproblems. Specifically, subproblem $i$, which corresponds to subcarrier $k_0$ in cell $n_0$ is formulated as:
\[
\max_{\mathbf{A}, \mathbf{B}} R_{LB,i}(\mathbf{P}^m, \mathbf{A}_i, \mathbf{B}_i)
\]
\[
s.t. \quad C1' : \sum_u a^k_{u,n_0} + \sum_u \sum_j b^k_{u,j,n_0} \leq 1,
\]
\[
C2' : a^k_{u,n_0} \in \{0, 1\}, b^k_{u,j,n_0} \in \{0, 1\}, \forall u, j,
\]
where $\mathbf{A}_i = \{a^k_{u,n}|n=n_0, k=k_0\}$, $\mathbf{B}_i = \{b^k_{u,j,n}|n=n_0, k=k_0\}$ and $R_{LB,i}$ is given by
\[
\sum_u a^k_{u,n_0} R^{k_0}_{u,n_0,1}(\mathbf{P}^m) + \sum_u \sum_j b^k_{u,j,n_0} R^{k_0}_{u,j,n_0,2,LB}(\mathbf{P}^m).
\]
Note that this subproblem can easily be solved by searching the optimal mode, MS and RS which provide the maximum rate for subcarrier $k_0$ in cell $n_0$ with $\hat{\mathbf{P}} = \mathbf{P}^m$.

C. PA Optimization

In this subsection, the PA stage for iteration $m$ is considered. After setting the indicators to $\{\mathbf{A}^{m-1}, \mathbf{B}^{m-1}\}$, the transmission modes are fixed in all cells. We denote by $\mathbf{S}_{u,n}(d)$ the set of subcarriers allocated to MS $d_{u,n}$ in direct mode and $\mathbf{S}_{u,j,n}(r)$ the set of subcarriers allocated to MS $d_{u,n}$ and RS $r_{j,n}$ in relay aided mode. Then the objective function of problem (5) can be rewritten as $R(\mathbf{P}, \mathbf{A}^{m-1}, \mathbf{B}^{m-1})$, given by:
\[
\sum_n \left( \sum_{k \in \mathbf{S}_{u,n}(d)} \log (1 + \Gamma^k_{d_{u,n},t_1}) + \sum_{k \in \mathbf{S}_{u,j,n}(r)} \log \min \left( 1 + \Gamma^k_{r_{j,n},1}, 1 + \Gamma^k_{d_{u,n},t_2} \right) \right),
\]
(8)
Note that $R$ is a non-convex function due to the presence of interfering power terms in the denominators of $\Gamma_{d,a,n,t_1}$, $\Gamma_{t,j,n}$, and $\Gamma_{d,a,n,t_2}$. To solve problem (5), we first replace it with an minimization problem that is then solved by using the SC-LD algorithm.

By introducing slack variables $\Psi_k = \{\Psi_{nk}, \forall n, k \in S_{u,jn}(r)\}$, the equivalent minimization problem is obtained as:

\[
\min_{P, \Psi_r} \sum_n \left( \sum_{k \in S_{u,n}(d)} \log \frac{f_{d,a,n,t_1}^k}{g_{d,a,n,t_1}^k} + \sum_{k \in S_{a,jn}(r)} \Psi_{nk}^k \right)
\]

s.t. $C3$, $C4$,
\[
C5: \log \frac{f_{d,a,n,t_1}^k}{g_{d,a,n,t_1}^k} - \Psi_{nk}^k \leq 0, \forall n, k \in S_{u,jn}(r),
\)
\[
C6: \log \frac{f_{d,a,n,t_2}^k}{g_{d,a,n,t_2}^k} - \Psi_{nk}^k \leq 0, \forall n, k \in S_{u,jn}(r),
\]

where $g_{d,a,n,t_1}^k = f_{d,a,n,t_1}^k + P_{s,a} G_{s,a,n}^k$, $g_{d,a,n,t_2}^k = f_{d,a,n,t_2}^k + P_{s,a} G_{s,a,n}^k$ and $g_{d,a,n,t_1}^k = f_{d,a,n,t_1}^k + P_{s,a} G_{s,a,n}^k$.

Problem (9) is still non-convex. In order to solve it, we now propose the SC-LD algorithm. To derive it, two steps are involved. During the first step, a convex approximations of problem (9) is constructed. We first use the method of Lemma 1 in paper [8] to condense all denominator posynomials $\{g_{d,a,n,t_1}^k, g_{d,a,n,t_2}^k, g_{d,a,n}^k\}$ into monomials $\{g_{d,a,n,t_1}^k, g_{d,a,n,t_2}^k, g_{d,a,n}^k\}$ using tentative PA results $P_0 = \{P_{s,a}, 0, P_{s,a}, 0\}$. Then, by introducing the logarithmic change of variables (e.g. $\tilde{x} = \log x$), the approximated problem is formulated as:

\[
\text{minimize } \sum_n R^n(\tilde{P}, \tilde{\Psi}_r)
\]

s.t. $C3$, $C4$,
\[
C4: \frac{1}{P_T} \sum_k (e^{\tilde{f}_{d,n,t_1}^k} + e^{\tilde{f}_{d,n,t_2}^k}) \leq 1, \forall n,
\)
\[
C5: \log \frac{f_{d,a,n,t_1}^k}{g_{d,a,n,t_1}^k} (e^{\tilde{P}}) e^{\tilde{\Psi}_r} \leq 0, \forall n, k \in S_{u,jn}(r),
\)
\[
C6: \log \frac{f_{d,a,n,t_2}^k}{g_{d,a,n,t_2}^k} (e^{\tilde{P}}) e^{\tilde{\Psi}_r} \leq 0, \forall n, k \in S_{u,jn}(r),
\]

where $R^n(\tilde{P}, \tilde{\Psi}_r)$ is given by:

\[
\sum_{k \in S_{u,n}(d)} \log \frac{f_{d,a,n,t_1}^k}{g_{d,a,n,t_1}^k} (e^{\tilde{P}}) + \sum_{k \in S_{a,jn}(r)} \log e^{\tilde{\Psi}_r}.
\]

$\tilde{P}$ denotes the vector of logarithmic power variables, $\tilde{\Psi}_r$ denotes the vector of slack variables $\tilde{\Psi}_{nk}$. Note that the method we used in this step belongs to the class of successive convex approximation methods [9]. After some mathematical calculations, it is easy to verify that all the approximations of problem (9) satisfy the three conditions proposed in [9] for the convergence of the successive approximation method.

During the second step, the Lagrange dual method is applied to solve the approximated convex problem. Let’s denote the Lagrange multipliers of $C4$, $C5$ and $C6$ by $\lambda = \{\lambda_n\}$, $\mu = \{\mu_{nk}\}_{k \in S_{u,jn}(r)}$ and $\nu = \{\nu_{nk}\}_{k \in S_{a,jn}(r)}$ respectively. Then the dual problem of (10) is written as $\max_{\lambda, \mu, \nu} \min_{\tilde{P}, \tilde{\Psi}_r} L(\lambda, \mu, \nu, \tilde{P}, \tilde{\Psi}_r)$. Here $L(\lambda, \mu, \nu, \tilde{P}, \tilde{\Psi}_r)$ denotes the Lagrange of problem (10).

After taking the derivative of $L$ w.r.t. $\tilde{\Psi}_{nk}$ and setting it to zero, we obtain $\mu_{nk} + \nu_{nk} = 1, \forall n, k \in S_{u,jn}(r)$. By applying it, the dual variables are updated with the subgradient method as:

\[
\lambda_n^{(i+1)} = \lambda_n^{(i)} + \delta_\lambda^{(i)} \left( \frac{1}{P_T} \sum_k (P_{s,a}^{(i)} + P_{\nu}^{(i)}) - 1 \right),
\]

\[
\mu_{nk}^{(i+1)} = \mu_{nk}^{(i)} + \delta_\mu^{(i)} \log \frac{f_{d,a,n,t_1}^k}{g_{d,a,n,t_1}^k} (P_{\nu}^{(i)}) e^{\tilde{\Psi}_r}.
\]

\[
\nu_{nk}^{(i+1)} = 1 - \mu_{nk}^{(i+1)}.
\]
Finally, the SC-LD algorithm is described in algorithm 1.

**Algorithm 1 SC-LD Algorithm**

1: **Input:** RA results of outer iteration \( m - 1 \), i.e. \( P^{m-1} \), \( A^{m-1} \) and \( B^{m-1} \);
2: **Initialize:** set \( t = 0 \), \( P_0 = P^{m-1} \), initialize \( S_{ujn}(r) \) and \( S_{mn}(d) \) using \( A^{m-1} \) and \( B^{m-1} \);
3: repeat
4:   **Initialize:** set \( i = 0 \), \( \mu_k^{(i)} = \mu_0 \), \( \forall n, k \in S_{ujn}(r) \), \( \lambda_n^{(i)} = \lambda_0 \), \( \forall n, k \) and \( P^{(i)} = P_0 \);
5:   repeat
6:     \( i = i + 1 \);
7: **Update power:** Update power variables \( P^{(i)} \) using equations (16) and (17);
8: **Update Lagrange factors:** Update Lagrange factors \( \{ \lambda^{(i)}, \mu^{(i)}, \nu^{(i)} \} \) using equations (11), (12) and (13);
9:   **until** \( ||\lambda^{(i)} - \lambda^{(i-1)}|| \leq \varepsilon_2 \)
10: \( t = t + 1 \);
11: Assign the results \( P^{(i)} \) to \( P_0 \);
12: **until** \( ||R^{m,t} - R^{m,t-1}|| \leq \varepsilon_3 \)
13: **Output:** \( P^m = P_0 \).

**IV. NUMERICAL EXPERIMENTS**

In this section, we evaluate both the convergence and the effectiveness of our algorithm. To illustrate the convergence, we first present sum rates w.r.t. iterations for one particular channel realization. Then results averaged over many channel realizations are provided and discussed to illustrate the effectiveness of our proposed algorithm.

**A. System setup**

As shown in Figure 1, we consider a multi-cell OFDMA system with \( N = 19 \) coordinated cells and \( K = 32 \) available subcarriers. Each cell contains \( U = 4 \) MSs and \( R = 3 \) RSs. Red squares, blue diamonds and black dots denote the BSs, RSs and MSs, respectively. The channel impulse response (CIR) of each link is randomly generated using a 8-tap delay line model. Each tap obeys a circular complex gaussian distribution with zero mean and variance as \( \sigma^2 \). Here \( \frac{\sigma_r^2}{\sigma_i^2} = \varepsilon^3 \) and \( \sum_{i} \sigma^2 = d^{-3} \). We set \( P^0_F = P_F, \forall n, \sigma^2 = -80dBm, \lambda_0 = K, \mu_0 = 0.5, \alpha_x = K/4, \alpha_y = 0.02, \varepsilon_1 = R_{ini}/100, \varepsilon_2 = 1 \) and \( \varepsilon_3 = R_{ini}/100 \). Here \( R_{ini} \) denotes the sum rate calculated with the initial RA results at the beginning of each iteration.
B. Convergence Illustration

In order to illustrate the convergence of our algorithm, a set of channels are randomly generated with $N = 7$ and $P_T = 50$ dBm. As shown in Figure 2, the total rate keeps increasing continuously and converges smoothly to the final rate. After convergence, the total rate is increased by around 18% and 15% using the CA in [6] and the proposed low-complexity algorithm (LCA), respectively.

C. Comparison of Average Performances

In order to illustrate the effectiveness of the proposed LCA, a series of system setups are randomly generated with different cell numbers ranging from 2 to 18 and $P_T = 30$ dBm, $\forall n$. For each system setup, average execution times and performances of the CA in [6] and the LCA are presented over 100 random realizations of channels. For comparison, we introduce two benchmark algorithms (BAs). Here BA1 corresponds to the BASE scheme considered in [5], when the CCI is always set to 0. BA 2 corresponds to the algorithm when the power is uniformly allocated and the RS and MS are randomly selected for each subcarrier.

As shown in Figure 3, both the CA and the LCA outperform the BAs, especially when the cell number is quite large. The LCA result in sum rates similar to those obtained with the CA, while its execution time is much less than that of the centralized one. Thus the LCA provides good tradeoff between performance and complexity, meaning that it is much more practical than the centralized one.

V. CONCLUSION

We have considered a multi-cell OFDMA downlink system with several DF RSs aiding the BS transmissions. The considered RA problem has been formulated in a delicate way to facilitate the algorithm
design. An iterative low-complexity algorithm has been proposed to solve the formulated problem. Through numerical experiments, we have illustrated the convergence of the LCA as well as its benefits compared with a CA. In future, subcarrier pairing and a higher efficiency protocol will be investigated, where the BS would be allowed to transmit a new symbol during the second TS.

VI. ACKNOWLEDGMENTS

The authors would like to thank Prof. S. Boyd for sharing the GGPLAB toolbox online, China Scholarship Council (CSC) for funding this work and BELSPO for the support of the IAP BESTCOM network. The work of T. Wang is supported by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

REFERENCES

Ad Hoc Voting on Mobile Devices

Manu Drijvers, Pedro Luz, Gergely Alpár and Wouter Lueks
Institute for Computing and Information Sciences (iCIS),
Radboud University Nijmegen, The Netherlands.

May 20, 2013

Abstract. We describe an ad hoc voting scheme that allows any group of people to very easily cast electronic votes in ad hoc settings. One of the challenges of online voting schemes is to ensure security and privacy. We propose a new approach, the ad hoc voting pattern, that meets these requirements. To show the feasibility of the proposed solution we implemented such a scheme on the Android platform. Moreover, due to the access point functionality of modern smartphones and the fact that the proposed scheme runs completely on an ad hoc fashion, the implemented scheme does not require internet access.

1 Introduction

The last decade has seen a clear shift from traditional mass communication to modern and personalized interactions. Mobile devices and ad hoc relations play an important role in this shift. As such, new protocols and schemes have been designed in order to provide secure, user-friendly and robust applications. An example of such applications is electronic voting, also known as e-voting, designed to provide an alternative and fast way of voting.

Electronic voting schemes must provide the same security properties as traditional paper ballot voting. The security requirements of such protocols include integrity, voter authentication, and ballot secrecy. These represent a big challenge for ad hoc networks since it is easy to eavesdrop on connections or tamper with protocols by connecting extra devices wirelessly. Integrity is one of the most important requirements for a voting protocol: no one should be able to tamper with votes. Also ballot secrecy is very important, as it enables sincere voting, by mitigating external factors such as coercion, social pressure, intimidation and bribery.
An electronic voting scheme following the ad hoc voting pattern can be thought of as the following three-stage protocol: 1. Acquire an ephemeral identity using a blind signature scheme (Figure 1a); 2. Vote using this identity and distribute the secret ballot using secret sharing (Figures 1b and 1c); and 3. Combine all secret fragments to calculate the result (Figure 1d).

Our work covers the analysis of existing e-voting schemes present in the literature, and a new general ad hoc voting pattern, inspired by Parakh and Kak [11]. We also propose a new scheme, that instantiates this pattern.

Our paper is structured as follows. First, we introduce some of the existing e-voting schemes in Section 2. Next, we generalize Parakh and Kak’s [11] proposal to a three-stage pattern, Section 3. We demonstrate the feasibility of the proposed scheme by

Fig. 1: Ad Hoc Voting Pattern - (a) Registration stage, (b)+(c) Voting stage, (d) Counting stage.
an Android implementation described briefly in Section 4. We conclude the paper in Section 5.

1.1 Related Work

Kiayias and Yung [10] define requirements for boardroom elections and a protocol that matches these requirements. This is first improved upon by Groth [8] and then by Hao et al [9]. However, even their scheme [9] becomes inefficient for elections with many candidates, as the computational complexity of tallying grows exponentially in the number of candidates.

Helios [1] is a web-based implementation of an e-voting scheme based on Benaloh’s simple verifiable elections [4]. This implementation has shown to be successful [2]. A downside of Helios is that it needs an external voting server to host the election.

All of the voting schemes mentioned above use some form of ElGamal encryption [7], due to its homomorphic properties, and zero-knowledge proofs [6]. The latter are typically used to show that ballots were formed correctly.

Parakh and Kak proposed a voting scheme based on implicit data security [11]. The scheme has security flaws of which the most devastating is that anyone can vote multiple times because of a broken signature scheme. Despite its weaknesses, their work inspired the ad hoc voting pattern proposed in this paper.

2 Technical Background

There are several cryptographic primitives that allow one to create secure e-voting schemes. In this section we provide a description of two different primitives: Shamir’s secret sharing scheme and Schnorr’s blind signature scheme.

2.1 Notation

In the remainder of this paper we will use the following notation. We write $a \in_R A$ to denote that $a$ is chosen uniformly at random from the set $A$. Furthermore, we write $\mathbb{Z}_q$ for the group consisting of the integers modulo a prime $q$.

2.2 Shamir’s Secret Sharing Scheme

Shamir’s secret sharing scheme can be used to make $t$-out-of-$n$ sharing of a secret $s$. Such a sharing consists of $n$ shares, one for each party. The secret $s$ can only be recovered when at least $t$ parties combine their shares. The mathematical principle behind Shamir’s secret sharing is that to reconstruct a polynomial of degree $t-1$, one needs at least $t$ points on this polynomial. By encoding the secret in the polynomial, usually as the constant coefficient, one obtains a secret sharing scheme.

2.2.1 Distribution More precisely, the Shamir’s Secret Sharing Distribution algorithm $SSSD(s, t, n)$ to create a $t$-out-of-$n$ secret sharing of a secret $s \in \mathbb{Z}_q$ consists of the following steps.

1. Choose coefficients $a_1, \ldots, a_{t-1} \in_R \mathbb{Z}_q$ and define the degree $t-1$ polynomial by $f(x) = s + a_1 x + \ldots + a_{t-1} x^{t-1}$.
2. Create the $n$ secret shares $s_i$ by setting $s_i = f(i)$. 
2.2.2 Reconstruction  Given a subset of \( t \) shares \( s_i, i \in I \) it is possible to reconstruct the polynomial \( f \) and hence recover the secret \( s \). Generally, the polynomial is reconstructed using Lagrange polynomials, but since we are only interested in \( f(0) \) we only need the following Lagrange coefficients:

\[
I_j^I = \prod_{i \in I \setminus \{j\}} \frac{i}{i - j}
\]

Then the secret \( s \) can be recovered by setting \( s = f(0) = \sum_{i \in I} I_j^I s_i \).

2.3 Schnorr Blind Signature Scheme

In traditional digital signatures the signer knows the message it is going to sign. However, in a blind signature scheme the signer can sign a message without knowing that message. We use this in our voting pattern because this allows us to decouple registration, where you are identifiable, from voting, where you want to be anonymous.

The first example of a blind signature scheme is described by Chaum [5]. In this paper we use Schnorr’s blind signature scheme. It is a variant of Schnorr’s identification protocol. Consider a signer with private key \( x \) and public key \( h = g^x \). In Figure 2 we show how this signer blindly signs a message \( m \) held by the recipient. In the following we write \((c', r') = SBSSign(x, m)\) to denote this interactive protocol. The blind signature can then be verified by checking whether \( c' \equiv H(g^c y^{-c'} \mod p|m) \). This scheme is provably secure if \( H \) is modelled as a random oracle [3].

| Signer | | Recipient |
|----------------------|----------------------|
| Secret: \( x \) | Message: \( m \) |

\[
w \in_R \mathbb{Z}_p^* \]
\[
a := g^w \mod p \]
\[
\alpha, \beta \in_R \mathbb{Z}_q \]
\[
a' := a \cdot g^\alpha y^\beta \mod p \]
\[
c := H(a' \| m) \]
\[
\beta \in_R \mathbb{Z}_q \]
\[
c' := \beta \mod q \]
\[
r := c \cdot x + w \mod q \]
\[
r' := r + \alpha \mod q \]
\[
a' := g^c \cdot y^{-c'} \mod p \]
\[
\text{Signature: } (c', r') \]

Fig. 2: Messages exchanged between Signer, with private key \( x \) and public key \( y = g^x \) and Recipient in order create a blind signature of the message \( m \). The values \( p, q, g \) are system parameters [1] and \( H \) is a cryptographic hash function mapping strings onto \( \mathbb{Z}_q \).

2.4 Voting scheme properties

Despite the variety of e-voting schemes that exists, most try to achieve the same set of security properties. The security properties for electronic vote can be categorized in two groups: correctness requirements and privacy requirements. A reliable scheme must provide the same level of correctness as the traditional paper ballot voting. The idea of reliable systems is based on the accuracy of the scheme, its integrity and democracy.

---

1 An instantiation of such parameters can be found in Section 3.2.1
Fig. 3: The Ad Hoc Voting Pattern in 8 steps. The registration authority (RA) registers eligible voters, the online polling booth (OPB) enables voting and tallies the votes afterwards, and the voting servers (VS) register the votes.

On the other hand, a system can be considered as privacy-friendly if it assures ballot secrecy and anonymity.

Schneier [12], introduced these security properties in the scope of the electronic voting by defining the following requirements: 

- **S1**: Only authorized voters can vote and at most once (eligibility).
- **S2**: No one can determine what anyone else voted (ballot secrecy).
- **S3**: No one can change anyone else’s vote without being discovered (integrity).
- **S4**: No one can duplicate anyone else’s vote (integrity).
- **S5**: Each voter can verify that his/her vote was counted (verifiability).

### 3 Ad Hoc Voting scheme

The proposed solution is a voting scheme that operates in an ad hoc scenario. In this section, we start by providing an overview of the ad hoc voting pattern, followed by an instantiation. Finally, we analyze the security of the scheme.

The system contains four types of parties: voters, one registration authority (RA), one online polling booth (OPB) and $n$ voting servers (VS). In the following we assume that we always require at least $t$ voting servers to provide final results. For an ad hoc voting scheme, we propose to have every voter to simultaneously act as a voting server.

#### 3.1 Ad hoc voting pattern

The proposed voting scheme can be described as a three stage protocol. The steps of each stage are depicted in Figure 3.

The first stage is the registration stage. First the voter authenticates him- or herself to the registration authority (step 1). If the voter is eligible, the RA will blindly sign a random identity generated by the voter (step 2).
The second stage is the voting stage. The online polling booth (OPB) will verify the blind signature to determine eligibility (step 3). The blindness of the signature ensures that the OPB and the RA cannot collude to recover the identity of the voter. Only the random identity is known to the OPB. Also, the voter will send a key to the OPB (step 3), which will then be forwarded to the voting servers (VS) (step 4), to allow the voter to access the VSs. Finally, the voter creates its ballot and creates $n$ shares using a $t$-out-of-$n$ secret sharing scheme (step 5). It uses its key to send these shares securely to each of the VSs (step 6).

When all votes have been cast the OPB closes, and initiates the counting phase. The OPB collects at least $t$ shares of each ballot and reconstructs the original vote (step 7). Each ballot is counted and the final result is published (step 8). In the end, the secret shares are published allowing the voters to verify that their ballot was correctly counted.

### 3.2 Ad hoc voting scheme

After a high-level overview over the ad hoc voting pattern, we provide the detailed description of scheme, where we use explicit cryptographic primitives. We start by setting up the system parameters necessary for the cryptographic primitives and then we proceed with the description of each stage.

#### 3.2.1 Set up

To setup the system, first generate two primes $p$ and $q$ of 1024 and 160 bits respectively, such that $q$ divides $p - 1$. Next, fix a generator $g \mod p$, such that $g$ has order $q$ in $\mathbb{Z}_p$, and publish the parameters $p, q$ and $g$. Next, the RA generates a private key $x \in R \mathbb{Z}_q$ and publishes its public key $y = g^x \mod p$. Furthermore, each of the voting servers generates a private key $u_i$ and publishes its public key $h_i = g^{u_i} \mod p$.

#### 3.2.2 Registration stage

In the registration stage the voter authenticates itself with the RA and obtains a blind signature on a random identity that will be used to cast the vote. Concretely this can be implemented as follows.

1. The voter authenticates itself to the RA and sends it a random identity $r_{id} = g^r$ with $r \in R \mathbb{Z}_q$.
2. The RA verifies the eligibility of the voter, and aborts if the voter is not eligible.
3. The voter and the RA simultaneously execute the $\text{SBSSign}(x, r_{id})$ protocol. In the end the voter has a signature $(c, r)$ on its random identity.

#### 3.2.3 Voting stage

At the beginning of the voting stage the voter creates an ephemeral private key $v_i$ and sets its public key to $V_i = g^{v_i} \mod p$. It sends its random identity $r_{id}$, the signature $(c, r)$ on that identity, and its public key to the OPB. If the signature is valid, the OBP forwards the public key to the VSs to allow the identity $r_{id}$ to cast a vote. The VSs only accept these messages from the OBP.

To cast its ballot $b$, the voter creates Shamir secret shares $(b_1, \ldots, b_n) = \text{SSSD}(b, t, n)$ of this ballot. Then, it calculates a shared key $k_i = h_i^{u_i}$ with voting server $i$. First, the voter identifies itself to voting server $i$ by sending $r_{id}$, and then it sends the share $b_i$ encrypted using $k_i$. The voting server can derive the same key by setting $k_i = V_i^{u_i}$, and thus can store the vote $(b_i, r_{id})$. 


3.2.4 Counting stage In this phase, the OPB collects all the ballot pieces from \( t \) VSSs and reconstructs the ballots using Shamir’s secret sharing reconstruction scheme. Now, the OPB simply tallies the ballots and publishes the result together with the polynomials (indexed by \( r_{id} \)). This allows every user to verify that his vote was correctly counted.

3.3 Analysis of security requirements

In section 2.4 we presented a list of requirements for electronic voting schemes. We now show how our ad hoc voting scheme meets such requirements.

The requirement \( S1 \) is met because to cast a vote one needs a signed anonymous identity, and only the RA can produce these signatures. The RA will only sign an anonymous identity for eligible voters. Furthermore, the RA will only sign one anonymous identity for each voter, and with a single anonymous identity only one vote can be cast.

A voter only shows its true identity to the RA, over a secure channel. This means that the RA is the only one that knows a true identity. However, the RA never learns the anonymous identity of the voter, because it blindly signs it. All other steps always refer to this anonymous identity, hence the true identity remains hidden. Thus requirement \( S2 \) is satisfied.

At least \( t \) VSSs need to cooperate in order to change a ballot. The properties of the secret sharing scheme ensure that any smaller coalition can only randomize the ballot, which with only very low probability will give a valid ballot again. So requirement \( S3 \) is also satisfied.

As long as less than \( t \) VSSs are corrupted, duplicating someone else’s vote is impossible, \( S4 \). To duplicate a vote, one has to know the vote. This vote is distributed over all VSSs, and \( t \) shares are required to reconstruct the vote. If less than \( t \) VSSs are corrupted, the vote cannot be determined and can therefore not be copied.

At the end of the voting phase, the OPB publishes the final result, as well as a list of all the polynomials received. This allows every voter to verify whether his vote, identified by the polynomial, is included in the list (\( S5 \)). Furthermore, each voter can verify if all the reconstructed polynomials contain a valid \( f(0) \). This allows a voter to verify that the sharing and reconstruction algorithms for all polynomials were correctly executed and that no one tried to tamper individual shares, \( S3 \).

4 Implementation

We implemented the ad hoc voting scheme on the Android platform. One device will act as the host, and perform the tasks of the RA and OPB. Having one device carrying out both tasks is secure under the assumption that the communication channel does not reveal anything about the identity. This device is excluded from voting. Every voter will also act as VS. TCP connections are used for all the communication. In practice one should use transport layer security, which has been omitted for this proof-of-concept. The threshold used in Shamir’s Secret Sharing is equal to the number of voting servers, which means that, unless all voting servers are corrupted, ballots cannot be reconstructed. Our proof-of-concept uses open authentication, which means that anyone able to connect to the RA is able to vote. Since modern smartphones can
function as a Wi-Fi access point, an ad hoc network hosted by one of the smartphones can be used to perform the voting.

5 Conclusion

We have proposed the ad hoc voting pattern, a new approach to ad hoc voting based on anonymous identities and secret sharing. An instantiation of this pattern has been given, based on Schnorr’s blind signatures and Shamir’s secret sharing, which fulfils desired properties such as ballot secrecy, eligibility, integrity and verifiability. One can easily think of other instantiations using different cryptographic primitives. For future work it would be interesting to analyze different instantiations and analyze their security. This might allow us to make the security requirements posed on the blind signature scheme and the secret sharing scheme more explicit. One could also research different authentication models, for example using out-of-band channels like NFC. Finally, it would be interesting to make a fully secure implementation of the ad hoc voting scheme.

References

Abstract. For more than three decades, the Data Encryption Standard (DES) was one of the most widely used cryptographic algorithms. It is still the dominating block cipher for banking applications. The DES was designed by IBM, verified by NSA and published by the National Bureau of Standards as a US Federal Information Processing Standard (FIPS) in 1977. The algorithm itself was fully public but the complete design criteria were only revealed by Coppersmith in 1994. He states that the IBM team was aware of differential cryptanalysis; the DES S-boxes are chosen to satisfy eight design criteria in order to resist this powerful attack. In their 1982 book, Meyer and Matyas state that the DES S-boxes were chosen so that they can be implemented with a minimum number of logic circuits. They mention that for an early design, in which not all of the design criteria are satisfied, the number of minterms varies between 40 and 48. However, for the final design the number of minterms is either 52 or 53, which is the smallest possible number that satisfies all the design criteria. Our research attempts to validate the IBM claims by generating a large number of candidate DES S-boxes satisfying specific criteria and by evaluating their number of minterms.

Keywords: DES, S-box, minterm, differential cryptanalysis

1 Introduction

The publication of the DES algorithm in 1977 by the US National Bureau of Standards was surrounded by controversy. First, there were complaints that the key length of 56 bits was too short to resist exhaustive key search attacks by well-funded opponents [4]. Second, while the algorithm details were published, only part of the design criteria were revealed, which lead to the speculation that DES contains a hidden security weakness or trapdoor. In 1993, Wiener showed an effective design for a US$ 1 million machine that can recover a DES key in 3.5 hours [6], which makes it plausible that the US government could recover DES keys by exhaustive search in the late 1970s. In 1989, Biham and Shamir [1] demonstrated an attack against DES in 1989 using a technique called differential cryptanalysis; in response IBM claimed that this attack was known to the designers of DES and that the design criteria for the DES S-boxes contributed to the defense against this technique. In 1994 Coppersmith, one of the designers of the DES, presented a list of eight design criteria for the S-boxes [2], claiming that these criteria were used for the creation of the eight original DES S-boxes. This seems to settle the question about the trapdoors. In spite of this, the role of the NSA in the design of DES and in particular its S-boxes is not fully clear, as the public statements from IBM and NSA on this matter seem to be not fully consistent.

In 1982 Meyer and Matyas, two other members of the DES design team at IBM, discuss the implementation of the S-boxes and more specifically the number of minterms necessary to implement them [5]. They claim that an early design produced S-boxes with a number of minterms between 40 and 48. As more design criteria were added, this distribution shifted to the range from 52 to 59. The left tail of the distribution was chosen in order to minimize
the size of the implementation of the DES in hardware. The goal of this paper is to verify this claim in the hope that this can throw some light on the generation of the DES S-boxes.

First we generated a large number of S-boxes that satisfy the design criteria of IBM published in 1994 [2]; this turned out to be more difficult than expected. We also generated weaker S-boxes. Next we compute the distribution of the number of minterms of all these S-boxes in order to validate the claim by Meyer and Matyas.

This paper is organized as follows. Section 2 discusses the design criteria of the DES S-boxes made public by Coppersmith and lists the distributions provided by Meyer and Matyas. In Section 3 our methods to find strong S-boxes are described, followed by a discussion of our results in Section 4; this discussion includes a comparison to the claims of Meyer and Matyas. Section 5 concludes the paper.

2 Properties of DES S-boxes

2.1 Design Criteria

In [2] the following design criteria for the DES S-boxes are listed:

(S-1) Each S-box has six input bits and four output bits.
(S-2) No output bit of an S-box should be too close to a linear function of the input bits. (That is, if we select any output bit position and any subset of the six input bit positions, the fraction of inputs for which this output bit equals the XOR of these input bits should not be close to 0 or 1, but rather should be near 1/2.)
(S-3) If we fix the leftmost and rightmost input bits of the S-box and vary the four middle bits, each possible 4-bit output is attained exactly once as the middle four input bits range over their 16 possibilities.
(S-4) If two inputs to an S-box differ in exactly one bit, the outputs must differ in at least two bits. (If $|\Delta I_{i,j}| = 1$, then $|\Delta O_{i,j}| \geq 2$, where $|x|$ is the number of 1-bits in the quantity $x$.)
(S-5) If two inputs to an S-box differ in the two middle bits exactly, the outputs must differ in at least two bits. (If $\Delta I_{i,j} = 001100$, then $|\Delta O_{i,j}| \geq 2$.)
(S-6) If two inputs to an S-box differ in their first two bits and are identical in their last two bits, the two outputs must not be the same. (If $\Delta I_{i,j} = 11xy00$, where $x$ and $y$ are arbitrary bits, then $\Delta O_{i,j} \neq 0$.)
(S-7) For any nonzero 6-bit difference between inputs, $\Delta I_{i,j}$, no more than eight of the 32 pairs of inputs exhibiting $\Delta I_{i,j}$ may result in the same output difference $\Delta O_{i,j}$.
(S-8) Similar to (S-7), but with stronger restrictions in the case $\Delta O_{i,j} = 0$, for the case of three active S-boxes on round $i$.

Each 6×4 S-box can be split into four 4×4 S-boxes (rows), where the leftmost and rightmost input bits of the large S-box are used to select one of the four smaller S-boxes. Therefore, we can make a distinction between criteria that are applicable to these smaller S-boxes and those that apply to the larger 6×4 S-boxes. Note that criterion (S-3) implies that each 4×4 S-box is a 4-bit permutation, which simplifies the search for S-boxes.
As mentioned earlier, the leftmost and rightmost input bits of a 6×4 S-box select one of the 4×4 S-boxes for which the four middle bits are the input. Since only the two middle input bits are varied in (S-5), this criterion can be completely verified for 4×4 S-boxes. If all 4×4 S-boxes that are used to generate a 6×4 S-box are verified, then the 6×4 S-box will always satisfy this criterion. Criteria (S-2) and (S-4) cannot completely validate a 4×4 S-box, but they can already be used to eliminate permutations, that will certainly lead to invalid S-boxes.

In order to find valid 6×4 S-boxes, we first create 4-bit permutations and validate them using criteria from (S-2) to (S-5). Next, these permutations can be combined to create S-boxes and tested with criteria (S-2), (S-4) and (S-6) to (S-8).

The following lemma shows that any S-box that satisfies the above criteria leads to three other solutions.

**Lemma 1.** Let (0, 1, 2, 3) be the rows of a valid S-box that satisfies all eight criteria. Then the S-boxes with the row orderings (1, 0, 3, 2), (2, 3, 0, 1) and (3, 2, 1, 0) are also valid S-boxes.

**Proof.** By criterion (S-4), if two inputs to an S-box differ in exactly one bit, the outputs must differ in at least two bits. Consider any column on row i. Then changing any one bit in the input will result either in a different column on the same row, or in the same column on row i ⊕ 1 or i ⊕ 2. It will never result in the same column on row i ⊕ 3 since that would require an input difference of two bits. A rotation over all rows s.t. row i’ = i ⊕ j, where j ∈ {0, 1, 2, 3} will still preserve this property. □

To make the Lemma 1 more clear, consider the first DES S-box S1:

<table>
<thead>
<tr>
<th>14</th>
<th>4</th>
<th>13</th>
<th>1</th>
<th>2</th>
<th>15</th>
<th>11</th>
<th>8</th>
<th>3</th>
<th>10</th>
<th>6</th>
<th>12</th>
<th>5</th>
<th>9</th>
<th>0</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>15</td>
<td>7</td>
<td>4</td>
<td>14</td>
<td>2</td>
<td>13</td>
<td>1</td>
<td>10</td>
<td>6</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>14</td>
<td>8</td>
<td>13</td>
<td>6</td>
<td>2</td>
<td>11</td>
<td>15</td>
<td>12</td>
<td>9</td>
<td>7</td>
<td>3</td>
<td>10</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>12</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>9</td>
<td>1</td>
<td>7</td>
<td>5</td>
<td>11</td>
<td>3</td>
<td>14</td>
<td>10</td>
<td>0</td>
<td>6</td>
<td>13</td>
</tr>
</tbody>
</table>

Observe that S1(0) = 14 and S1(33) = S1(0b10001) = 15 which leads to a 2-bit input difference and 1-bit output difference. As we change the row ordering to (0,3,1,2), we get the following S-box S′₁:

<table>
<thead>
<tr>
<th>14</th>
<th>4</th>
<th>13</th>
<th>1</th>
<th>2</th>
<th>15</th>
<th>11</th>
<th>8</th>
<th>3</th>
<th>10</th>
<th>6</th>
<th>12</th>
<th>5</th>
<th>9</th>
<th>0</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>12</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>9</td>
<td>1</td>
<td>7</td>
<td>5</td>
<td>11</td>
<td>3</td>
<td>14</td>
<td>10</td>
<td>0</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>0</td>
<td>15</td>
<td>7</td>
<td>4</td>
<td>14</td>
<td>2</td>
<td>13</td>
<td>1</td>
<td>10</td>
<td>6</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>14</td>
<td>8</td>
<td>13</td>
<td>6</td>
<td>2</td>
<td>11</td>
<td>15</td>
<td>12</td>
<td>9</td>
<td>7</td>
<td>3</td>
<td>10</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Here S₁′(0) = 14 and S₁′(1) = 15. Therefore, with this S-box a 1-bit input difference can lead to a 1-bit output difference which contradicts criterion (S-4). However, any other rotation mentioned in Lemma 1 leads to a valid S-box.

This means that for four valid 4-permutations, we do not have to verify all the 4! combinations but only \( \frac{4!}{4} = 6 \).

### 2.2 The Number of Minterms

The 6×4 S-box is a multi-output Boolean function, that can be expressed in canonical form using minterms. The number of minterms in the canonical form is the number of logical
ports necessary to implement the S-box. Matyas and Meyer claim that this number can be used as a heuristic for cryptographic strength, because the number of minterms needed for the final design of S-boxes far exceeded the amount needed for an earlier design. We repeat their main results below; for more details the reader is referred to [5]. Table 1 (left) shows the distribution of minterms after reduction for the early design of S-boxes, based on a sample of 18 S-boxes. It is left open which criteria were already included in this design, but the authors do admit that this design was nearly random.

Table 1: Distribution of minterms for the early variant (left) and the DES S-boxes (right)

<table>
<thead>
<tr>
<th># Minterms per S-box</th>
<th># S-boxes</th>
<th># Minterms per S-box</th>
<th># S-boxes</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>1</td>
<td>52</td>
<td>3</td>
</tr>
<tr>
<td>41</td>
<td>1</td>
<td>53</td>
<td>7</td>
</tr>
<tr>
<td>44</td>
<td>3</td>
<td>54</td>
<td>9</td>
</tr>
<tr>
<td>45</td>
<td>3</td>
<td>55</td>
<td>22</td>
</tr>
<tr>
<td>46</td>
<td>4</td>
<td>56</td>
<td>16</td>
</tr>
<tr>
<td>47</td>
<td>2</td>
<td>57</td>
<td>20</td>
</tr>
<tr>
<td>48</td>
<td>4</td>
<td>58</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>59</td>
<td>2</td>
</tr>
</tbody>
</table>

As more design criteria were added, the number of minterms increased. The distribution for the final design of S-boxes is shown in Table 1 (right), based on a sample of 83 S-boxes. We assume that this design includes all criteria given by Coppersmith [2]. It is still unknown how many of these valid S-boxes exist. Matyas and Meyer used a sample of 83 S-boxes but the reason for this could simply be the limited computing capacity of the time.

According to Matyas and Meyer, the eight DES S-boxes were chosen from the left tail of this distribution (52 and 53 minterms) to make the implementation on a single chip as easy as possible.

In our research we have created a large amount of S-boxes using Coppersmith’s criteria and we have compared the distribution of their minterms to those given by Meyer and Matyas. We are interested to know which criteria were part of the early design and whether generating S-boxes according to the Coppersmith’s design criteria gives the minterm distribution of Table 1.

3 Finding Valid S-boxes

As described in Section 2.1, we focus our initial search on $4 \times 4$ S-boxes and then combine these to create valid $6 \times 4$ S-boxes. We first create random 4-bit permutations and evaluate them with criteria (S-2) to (S-5). Next these permutations were combined and tested with criteria (S-2), (S-4) and (S-6) to (S-8).

Our initial approach was to generate the $4 \times 4$ S-boxes randomly. However, as there are $16! \times \frac{4!}{4^4} = 125,536,739,328,000 \approx 2^{46.8}$ combinations of four permutations and only few valid
ones, random search turned out to be too slow to find a large number of valid permutations. Therefore, a more systematic method is necessary based on affine equivalence.

**Definition 1 (Affine Equivalence).** Let $A$ and $B$ be $n \times n$ invertible linear mappings over $GF(2)$ and $a$ and $b$ $n$-bit vectors over $GF(2)$. Then, the invertible S-boxes $S_1$ and $S_2$ are affine equivalent if the affine equivalence relation $S_1(x) = B^{-1} \cdot S_2(A \cdot x \oplus a) \oplus b$ holds. The set of such affine equivalent S-boxes are called an affine equivalence class.

In his PhD thesis, De Cannière [3, pp. 75–94] observes that for 4-bit permutations, 302 affine equivalence classes can be found. Since affine equivalence leads to the same algebraic and differential properties, it is more efficient to study one representative of each class. We hence tried to use permutations that are in the same affine equivalence classes as the permutations of the DES S-boxes. Table 2 shows to which affine equivalence classes the DES S-boxes belong with the numbering provided in [3].

<table>
<thead>
<tr>
<th>S-box</th>
<th>Equivalence class</th>
</tr>
</thead>
<tbody>
<tr>
<td>DES1</td>
<td>207 207 207 145</td>
</tr>
<tr>
<td>DES2</td>
<td>108 144 142 194</td>
</tr>
<tr>
<td>DES3</td>
<td>191 210 142 215</td>
</tr>
<tr>
<td>DES4</td>
<td>194 194 194 194</td>
</tr>
<tr>
<td>DES5</td>
<td>191 163 88 166</td>
</tr>
<tr>
<td>DES6</td>
<td>38 179 195 108</td>
</tr>
<tr>
<td>DES7</td>
<td>210 148 194 218</td>
</tr>
<tr>
<td>DES8</td>
<td>82 237 167 208</td>
</tr>
</tbody>
</table>

Since it is known that nonlinearity properties are unchanged by affine transformations, we can assume that all these 4-bit permutations satisfy criterion (S-2). The lists for each affine equivalence class were trimmed using criteria (S-4) and (S-5) and then the permutations were combined in the same way as in the corresponding DES S-box. For example the permutations from equivalence class 207 were used as first, second and third row in combination with a permutation from equivalence class 145. Finally, the invalid S-boxes were eliminated using criteria (S-2), (S-4) and (S-6) to (S-8). With this method, hundreds of valid S-boxes have been generated.

Finding valid S-boxes using the criteria in Coppersmith’s paper was much more computationally intensive than expected; we only managed to create valid S-boxes using the equivalence classes of the DES S-boxes. This leads to the conclusion that it is likely that the designers of the DES algorithm had more efficient algorithms to create strong S-boxes; one can even speculate that there may have been additional criteria to simplify the search.
4 Discussion on the Number of Minterms

4.1 The final design

After generating 635 S-boxes using the affine equivalence classes of the DES permutations, we used the Espresso logic minimizer\(^1\) to minimize their Boolean functions and calculate a distribution of the number of minterms necessary to implement them. Our distribution and that of Meyer and Matyas are shown in Figure 1; the hypothesis that the two distributions are the same is accepted with a significance of 9.15% by the Kolmogorov-Smirnov hypothesis test. However, when we calculate the number of minterms necessary to implement the 8 DES S-boxes, they appear to range from 52 to 55 and not from 52 to 53 as Matyas and Meyer suggested. The most plausible explanation for this small difference is that they have used a more powerful minimization tool.

\[\text{Fig. 1: Distribution of the number of minterms of S-boxes compared with the distribution of Meyer and Matyas}\]

4.2 The Early Design

Meyer and Matyas also provide a distribution of the number of minterms for an early design of the S-boxes, situated between 40 and 48 minterms. To identify the criteria for this early design, we repeat our generation process for S-boxes that satisfy fewer criteria. Leaving out criteria from (S-5) to (S-7) does not cause a significant change. When we leave out (S-4), the minimum number of minterms becomes 47, which still does not correspond to the distribution of Meyer and Matyas.

\[\text{Fig. 2: Distributions of the number of minterms of S-boxes for some of the criteria}\]

\(^1\) http://en.wikipedia.org/wiki/Espresso_heuristic_logic_minimizer
Only when the non-linearity criterion (S-2) is left out, a significant shift towards the left can be noticed, producing a distribution that corresponds to the Meyer and Matays distribution for the early design with a significance of 7.85%, again based on a Kolmogorov-Smirnov hypothesis test. However, the corresponding S-box is nearly random, since only criteria (S-1) and (S-3) are applied.

![Distribution of the number of minterms for S-boxes satisfying only criterion (S-1) and (S-3) compared to the distribution for the early design of Meyer and Matyas](image)

**Fig. 3: Distribution of the number of minterms for S-boxes satisfying only criterion (S-1) and (S-3) compared to the distribution for the early design of Meyer and Matyas**

## 5 Conclusion

We conclude that our search for S-boxes satisfying the eight criteria specified by Coppersmith [2] was much more computationally expensive than expected. This is rather surprising: according to Moore’s law the computational power for the same cost has increased by a factor 25 million between 1975 and 2012; even if one takes into account the fact that the budget of IBM may have been two or three orders of magnitude larger than ours and IBM probably has spent much more time optimizing their code, this still leaves a large gap. Even with this gap, we had to resort to S-boxes that start from the existing S-boxes designed by IBM. Hence one can conclude that the algorithm used by IBM to generate DES S-boxes was much better than the best publicly known algorithm today. This leaves the open problem whether we can come up with better algorithms and whether we can find S-boxes that are not related to the DES S-boxes. The minimum number of minterms for our S-boxes was a little larger than for the DES S-boxes; however, it seems unlikely that this difference is the consequence of an unpublished design criterion.

The early design of the S-boxes described by Meyer and Matyas in 1982 [5] likely corresponds to a random combination of 4-bit permutations. One can wonder whether such a variant has seriously been considered by IBM, since by today’s cryptographic standards such a DES variant is very weak.

## References

Signal Processing in Hearing Aids
Anne Hendrikse
GN ReSound,
et Eeuwsel 6, 5612 AS Eindhoven, the Netherlands
ahendrikse@gnresound.com

abstract
Classically, much of the hearing aid signal processing has focussed on the attenuation component of a hearing impairment, using amplification and compression to restore audibility. With the introduction of digital hearing aids several advanced signal processing strategies became available, which can aid speech understanding in noise, most likely caused by the distortion component of a hearing impairment. We review some of these signal processing strategies and relate them to knowledge on hearing impairment attenuation and distortion components and show that the problem is still far from solved. A big problem with these strategies is that they cause distortions themselves.

1 Introduction
People with a hearing impairment experience several problems as a result of that hearing impairment. One of the biggest complaints resulting from a hearing loss is a difficulty to hear and understand speech. A well known effect of a hearing impairment is that soft sounds are no longer perceived. This part of the hearing impairment can be treated by amplification, for which in the past many strategies have been developed.

However, a focus purely on the hearing loss is missing several aspects. Primary, hearing is not a linear relation between input power and perceived loudness. That relation has been measured in several studies, showing significant interpersonal variation. However, a general trend can be observed and this lead to the development of the iso226 standard [1] which can be represented by equal loudness contours, see figure 1a.

Associated with these equal loudness contours is the phon scale. Tones of different frequencies which are perceived as equally loud are assigned an equal phon. The phon of a tone is defined by the sound pressure level of a tone at 1 kHz with perceptually the same loudness.

The standard is both upper and lower limited. This reflects the workings of a normal hearing system: really soft sounds are not perceived by a normal hearing person either and extremely loud signals cause pain and before that level is reached, there is already a significant risk of damaging the hearing system.

A hearing impairment in general tends to increase the minimum sound pressure required before the sound is perceived. However, this change of hearing threshold is frequency dependent and the curves of a hearing impaired are therefore not simply an upwards shifted version of the curves in figure 1a.
Figure 1: (left) The curves indicate the sound pressure level for one phon for different frequencies. A phon of a tone is given by the sound pressure level of a tone at 1 kHz with perceptually the same loudness. (right) There is quiet some variation in hearing loss, but they are usually classified in two categories: ski slope, with a dominant high frequency loss and a flat hearing loss, where the loss is almost constant over all frequencies.

There is considerable variation in threshold alteration between hearing impaired. Therefore much of the diagnosis of a hearing impairment is focused on measuring these increased hearing thresholds. A classical approach is to present tones at varying levels and frequencies and ask which tones are perceived. From these measurements the increased hearing threshold can be estimated. The increase in threshold is usually expressed as a loss of hearing in dB and is graphically represented in an audiogram. Figure 1b shows audiograms of a few typical hearing losses [2].

Since the hearing loss is not uniform over the frequencies, many hearing aids support a frequency dependent amplification scheme. The settings of such a scheme are determined by a fitting rule, such as NAL-NL1 [3] with the audiogram as input. These rules also take into account that while a hearing impairment usually results in an increased hearing threshold, often the pain threshold has not been altered or even decreased, resulting in a loss of dynamic range. The pain threshold is however not determined during initial diagnosis. How the loudness perception between these two extremes is altered is also not known and still subject of study, including model generation. Still, to compensate for the loss in dynamic range, compression is part of the signal processing possibilities of many hearing aids.

As the loudness perception is not fully specified by the hearing thresholds, there seems to be room for improving the signal processing in the hearing aids. However, there are several studies indicating that the pure-tone average (PTA) hearing loss measurements do not capture all the relevant information of a hearing impairment. One of the biggest complaints of people with a hearing impairment is that they have difficulty with understanding speech in noisy situations. A significant number of hearing aid users do not perceive any benefit of their hearing aids in those conditions, while in some cases the hearing aids even seem to actually worsen the problems.

This problem has been the subject of many studies, but still there is no consensus on
a solution. Even the exact causes of the problems are still under investigation. In the following sections we review some of the studies on the speech in noise topic and review some of the strategies suggest in aiding hearing impaired with speech understanding.

2 Speech in noise

As already indicated in the introduction, while the hearing thresholds are an important part of the diagnosis of an hearing impairment, making sounds audible does not seem to solve the complaints associated with hearing impairment.

2.1 Reported issues with speech in noise

The problems with understanding speech in noise is demonstrated in consumer evaluation studies. In the MarkeTrak studies hearing aid users were interviewed on a number of aspects concerning their satisfaction with hearing aids [4].

One part of the study focussed on the user ratings of their hearing aids in specific situations. When asked about their satisfaction with their hearing aids, 91% is satisfied with their hearing aids in one-on-one situations, with 6% being dissatisfied. The one-on-one situation is characterised by a very limited amount of interference either from other speakers or other sounds. In contrast, only 68% is satisfied in large groups, with 20% being dissatisfied. This situation corresponds to the cocktail party phenomena and it shows that with interfering talkers, the dissatisfaction more than triples.

A previous study on why people no (longer) use their hearing aids, shows an even clearer picture. The number one reason reported was that the hearing aids provided poor benefit (29.6% of the respondents indicated this), already indicating the simply amplification strategy does not suffice to solve all hearing impairment relating problems. The second reason reported is because provides no benefit in background noise/noisy situations (25.3% of the respondents) [5].

In [6] it is claimed that most audiologists acknowledge that the most common complaint of adult patients about their hearing loss is their inability to understand speech in noise, especially multiple speech sources. This matches closely the reports of [5].

2.2 Hearing thresholds and noise

The complaints of the hearing aid users themselves and what is reported by audiologists indicate that the hearing impairment is not just a hearing loss. Several studies focus on the fact that the hearing thresholds are determined by PTA loss measured in quiet conditions, while the speech in noise problems occur under quite different conditions. These studies showed that the relation between hearing thresholds and speech in noise problems is rather weak.

[7] summarises a number of studies where the hearing loss is compared to the speech in noise performance. The hearing loss is measured in PTA loss in otherwise quiet conditions. The speech in noise performance was measured as signal to noise ratio (SNR) loss, defined as the increase in SNR required for a hearing impaired (HI) to recognize words with the same performance as normal hearing (NH). Although the
type of noise varied (for some tests it was a single speaker interference, while in others it was a 12-talker babble), it was reported that hearing thresholds is a poor indicator of speech understanding in noise. For hearing impaired with almost the same PTA loss, the SNR they required to obtain a 50% recognition rate varied approximately 15 dB.

In [8] the speech reception threshold (SRT) for sentences was investigated in quiet and under four noise conditions. Using 110 subjects (age 60-90) similar findings were reported for how accurate the speech hearing loss in noise can be predicted from the PTA measurements. It was shown that individuals with approximately the same hearing loss for speech in quiet conditions can vary considerably in their hearing loss for speech in noise conditions. Similar findings led Plomp to model the hearing impairment to have two components: an attenuation component and a distortion component [9].

The noise interference seems to be particularly bad when the interference is a (group of) talkers. This suggests that the dynamics are of importance here. In [10] an in depth study of human speech recognition is performed. One aspect that is taken into account is the manner in which this process is robust against noise. One particular strategy that is employed is to do the speech feature detection independently in separate bands. Models based on this assumption are then shown to match measurements of human speech recognition performance. This justifies the multi band signal processing approach taken in many hearing aids nowadays.

2.3 Frequency resolution

One of the effects a hearing impairment is suggested to have, besides the PTA loss, is that it can affect the frequency resolution of the hearing system. So if one of the strategies human speech recognition employs to make speech recognition more robust against noise is to do speech feature recognition independently in separate frequency bands as [10] suggested, then if the frequency resolution of the hearing system is affected, this could impair this spectrum based separation and thus be one of the factors underlying the problems with understanding speech in noise reported by hearing impaired. The frequency resolution depends on the selectivity structures in the cochlea. Several studies have attempted to measure this frequency resolution loss by measuring the auditory filter shapes.

In [11] the auditory filter shapes were studied for both ears with normal hearing and ears with a hearing impairment resulting from an almost pure cochlear impairment. The shapes were estimated for center frequencies of 0.5, 1 and 2 kHz. The procedure was to estimate the detection level of a target tone while maskers were notched at the target frequency. By varying the width of this notch, an estimate of the side band interference could be estimated.

Although much of the test was performed with only 6 subjects, the results clearly showed a significant reduction in selectivity of the impaired auditory systems in comparison to the normal hearing ears. For the normal ear, the -10 dB attenuation of the noise was achieved if the notch was roughly 0.2 to 0.3 times the center frequency tested. For impaired ears, at 0.5 kHz it turned out to be impossible to determine the selectivity for 3 of the 5 subjects, while for the fourth subject the required notch width had to be almost doubled.

At 1 and 2 kHz, the determined filters showed particularly a widening towards the
lower frequencies, where the notch width had to be doubled or more to get the same amount of attenuation. For some subjects also the notch width required on the side of the higher frequencies increased significantly, but for 3 of the 5 subject it remained of the same order or even decreased slightly.

Part of the study presented in [12] tested the frequency resolution in a fairly similar way, however, the subjects where not selected on the based on the origin of the hearing impairment, unlike the study reported in [11]. Their results showed no significant difference between NH and HI in equivalent rectangular bandwidth (ERB), a measure closely related to the notch width in the [11] study, for 0.5 and 1 kHz center frequencies, but for 2 and 4 kHz there was a significant difference in ERB.

These results suggest that frequency resolution reduction occurs for a significant part of the hearing impaired, but the biggest problems occur for a smaller group with a specific kind of hearing loss.

2.4 Other distortion factors

A loss of frequency resolution is not the only factor in the distortion component of hearing impairment. Some other factors are timing related. One of such factors is the loss of so called phase locking. In [13] the effect of hearing impairment on the use of temporal fine information for speech recognition is studied, showing that partly due to loss of phase locking this is also one factor in the distortion component.

2.5 Extended modelling and diagnosis

Based on similar findings on how weak the SNR loss corresponds to the loss in hearing threshold measured by PTA loss, several authors suggest to extend the classical hearing impairment models and use other diagnosis besides the audiogram.

In [7] it was proposed to extent the articulation index (AI)[14], to include a channel capacity term to incorporate the fact that some speech cues can never reach the brain due to distortions, thereby incorporating that audibility itself is not enough for speech understanding.

To diagnose the distortion component of the hearing impairment, in [15] a variation on the SRT is defined: the SNR-50. This is the SNR required for 50% correct recognition of words. The SNR loss is then the increase in SNR-50 required by a HI in comparison to a NH. In [16] it is advocated that such speech in noise tests should be used during the selection of hearing aids for HI as it is an important factor and for some hearing aid users even the only factor for which the device should provide benefit. After an extensive review of the relation between the attenuation and distortion model with respect to word recognition both in quiet and background (babble) noise, in [6] also the case is made that word recognition performance should be measured when determining the appropriate intervention strategy for HI.

2.6 DSP solutions beyond amplification

Due to the problems HI experience when listening in noise, several signal processing strategies have been proposed to increase SNR. If a hearing aid has multiple mi-
Microphones, then beam forming can be used to give different amplification to sounds originating from different directions.

Another proposal is spectral subtraction, where it is attempted to classify in which bands contain the target signal and reduce the other bands in amplification. Although such methods show an improvement in SNR, an improvement in speech in noise understanding is hard to find [17].

3 Hearing impairment and DSP effects on localisation

In the previous sections we showed that a part of the problem hearing impaired experience with speech in noise is thought to originate from a reduced frequency resolution. A short summation was given of possible signal processing strategies which at least in theory can assist HI with speech in noise understanding.

Besides speech understanding, the auditory system also has other functions, one of which is localising the origin of sound sources. Moreover, the cues for performing this task are also shown to assist in speech understanding in cocktail party like settings. Several studies have demonstrated that this localisation task is also affected by a hearing loss, but what is a bigger problem for hearing aid users, is that some of the signal processing solution presented in the previous section actually distort the cues in the signal required for this task.

In [18] an extensive study into the effect of hearing impairment and hearing aid configuration is reported. In the literature review it is shown that previous results were inconclusive about the relationship between the PTA loss and the performance in localisation: in [19] it was concluded that ”localization ... performance is ... not easily predicted on the basis of audiograms” [18], demonstrating again that audiograms do not quantify all the problems associated with hearing impairment. However, in [18] it was shown some predictions can be based on the audiogram: a high frequency loss affects median vertical plane localisation. Frontal horizontal plane localisation is enabled by access to low and mid frequencies containing interaural time and/or level differences. Since these frequencies are in general not strongly affected in hearing loss, these localisation abilities are only mildly distorted. Conduction losses did show an effect on this localisation task. So hearing impairment has an effect on the localisation task.

However, strategies to aid in speech understanding affect the localisation task as well. In [20] it was shown different hearing aid types affect the localisation performance and if the source itself is made louder, hearing impaired performed better without hearing aids. Moreover, in [21] it was shown that signal processing strategies such as compression, beam forming and spectral subtraction also reduce localisation performance.

3.1 Cochlear implants

In some cases of hearing impairment parts of the auditory system are affected to such an extend, that cochlear implants are used. Cochlear implants bypass the transduction part of the hearing system and electrically stimulates the auditory nerve.
This approach mainly focusses on restoring the power envelope in a selected number of frequency bands, neglecting a number of features. First of all, cochlear implants only have in the order of 15 electrodes, compared to the roughly 3000 to 4000 inner hair cells, suggesting a loss of frequency resolution. Furthermore the current of the electrodes is spread broadly, frustrating the stimulation of specific fibres. Also the timing of the pulses is not dependent on the time structure of the sound.

As a result cochlear implant users particularly suffer from the cocktail party problem: they can have an acceptable restoration of speech understanding in quite conditions, but their ability to separate a target speaker from the background is severely reduced and the associated problems can be considered an extreme case in comparison to hearing aid users.

Studying what is the factor that provides benefit to cochlear implant (CI) users with still some acoustic hearing left, in [22] it was found that a significant part was contributed by the fundamental frequency cue available in the residual acoustical hearing. In [23] CI users are also used to get a better understanding on mechanisms of vowel perception. In that case it was shown that the vowel enhancement effect is also present in CI users, suggesting that it is not an effect of the cochlea itself.

The issues with pitch perception are extreme in the case of cochlear implants, because the devices do not transmit some of the information that is used for pitch perception. However, as is demonstrated [24], some types of hearing impairment affect pitch perception as well. Therefore if some solution is found for cochlear implant users, then this might be extended to hearing aids as well to mitigate some of the issues of certain types of hearing loss.

Indeed, the vowel enhancement strategy has been attempted in hearing aids as well. And although measured in SNR voiced speech enhancement shows a benefit [25], no improvement in speech recognition has been determined [26].

4 Conclusion

The research done on hearing impairment has increased the understanding of the causes of the speech in noise problems and some approaches to aid in speech in noise understanding have been suggested. The different validation tests however show that still large problems remain for some hearing impaired. The review shows that process of understanding speech in noise is a complex system of which some important components may yet still be unknown.

One major problem in developing hearing aids is that while some aspects of the hearing impairment can be treated, these solutions can interfere with other functions of the hearing system as has been demonstrated for example for localisation. Advanced signal processing strategies such as compression, beam forming and spectral subtraction are shown to cause distortions in spatial cues. These factors put further constraints on solutions to some of the hearing impairment related issues such as the reduced frequency resolution.
References


Optimization by adaptive simulated annealing: Intelligent light control using an array of dimmable LEDs

Xin Wang*, Jean-Paul Linnartz*, and Tjalling Tjalkens*

*Eindhoven University of Technology,
Dept. Electrical Engineering, Signal Processing Systems Group
{Xin.Wang, J.P.Linnartz, T.J.Tjalkens}@tue.nl

Abstract

In the past few years, LEDs have become a particularly popular and attractive light source, due to their high energy efficiency and long life time. Moreover, LEDs offer superior flexibility in controlling the output light patterns. This paper addresses the inverse problem of how to render an appropriate illumination for multiple users by controlling the dimmings of a dimmable LED array. Each of the users has different task light preferences and they are at different locations of an indoor environment. To solve this optimization problem, we invoke adaptive simulated annealing. The trade-off curve between reducing energy consumption and increasing user comfort is presented at the end of this paper.

1 Introduction

As one of the most popular light sources, LEDs are widely used in many lighting applications. Compared with traditional light sources, advantages include high energy efficiency, small size, quick on/off time, long lifetime and dimmability. When arranged as an array, the dimmability of LED offers more degrees of freedom to control the output light pattern. However, in most current lighting applications, this opportunity is not fully utilized. It is not practical to offer a separate control button for every individual LED. Common solutions fall back to 1) all the LEDs are controlled by one switch; 2) the LED array is divided into several subsets and each is controlled its own switch. For a 3m by 3m office which has a 5 by 5 LED array on the ceiling and a user located in the center, the simulated illuminance contour of the two scenarios are illustrated in Figure (1a) and (1b).

The forward problem of computing the light pattern that can be achieved given intensity of each LED, i.e., a dimming level between 0 and 100% of its maximum output lumen, can be analyzed by several 3D design tools for arbitrary room layout. However, in this paper we focus on the inverse problem: we want to optimize the dimming of each LED in the array such that every user gets comfortable task illumination, while at the same time the overall energy consumption is conserved, as illustrated in Figure 1c. We assume that we have full knowledge of the environmental layout, the parameters of every LED, the location and preference of each user.

The rest of the paper is organized as follows: Section 2 presents the overview of the smart LED lighting control system; then in Section 3 we propose a solution based on adaptive simulated annealing to minimize the objective function; finally, Section 4 concludes this paper.
2 Overview of smart LED lighting control system

Consider an indoor environment with a deterministic (fixed and known) layout consisting of $N_{LED}$ LEDs and a known or possibly statistically varying number of $N_{User}$ users, with $N_{LED} \gg N_{User}$. The power of every LED is controlled by a central controller and modeled as a continuous dimming fraction $W = \{w_1, w_2, \ldots, w_{N_{LED}}\}^T, (0 \leq w_n \leq 1, n = 1, 2, \ldots, w_{N_{LED}})$ of its maximum output power $P = \{p_1, p_2, \ldots, p_{N_{LED}}\}^T$. A parameter $\gamma$ between 0 and 1 is used to model the trade-off between reducing the lighting energy consumption and increasing user comfort. As a result, the inverse optimization problem can be modeled as

$$\begin{align*}
\arg \min_W & \quad (1 - \gamma)\frac{P^T W}{P^T 1} - \gamma S_A(X) \\
\text{subject to} & \quad 0 \leq W \leq 1, \\
& \quad X \succeq I_O, \\
& \quad \theta_T \geq 0.7, \\
& \quad \theta_O \geq 0.5.
\end{align*}$$

In the objective function, the first term is normalized energy consumption and the second one describes user comfort $S_A()$ for output light pattern $X$. Both of the terms range from 0 to 1. The first constraint means the dimming of every LED should be between 0 and 1; the second constraint means the output light pattern should be larger than or equal to the minimum (recommended) illuminance requirement $I_O$; the other two constraints ensure that the illuminance is evenly distributed. The following sections elaborate the objective function and the constraints.

2.1 Energy consumption

The energy consumption $E_{Array}$ of the lighting system can be modeled as the sum of every LED’s maximum power weighted by its dimming, i.e., the energy loss in the lamps is ignored, so

$$E_{Array} = P^T W.$$  \hfill (2)
calculated by the Generalized Lambertian Model [1]

\[ I = \frac{(m + 1)I_0}{2\pi h^2} \left( 1 + \frac{d^2}{h^2} \right)^{-\frac{m+1}{2}}, \]

where \( I_0 \) is total luminous flux (in Lumen) of the LED; \( d \) and \( h \) are horizontal and vertical distance from the LED to the point, expressed in meter. The Lambertian mode number \( m \) is a positive number which describes the width of the light beam. For an array of dimmable LEDs, as shown in (4), the illuminance of a certain point is the superposition of maximum illuminance can be provided by every LED in the array weighted by its dimming

\[ x = ITW, \]

where \( I = \{i_1, i_2, \ldots, i_{N_{LED}}\}^T \) is the illuminance vector of the LED array; \( i_n \) is the maximum illuminance that LED \( n \) could provide at this point, which can be derived from (3).

2.2 User comfort

The user comfort can be quantified by his or her satisfaction function [2], which ranges from 0 to 1 and describes the user comfort as a function of illuminance. Larger value of satisfaction function represents higher user comfort. An example satisfaction function is shown in (5)

\[ S(x) = \alpha e^{-\frac{(\ln x - \ln \xi)^2}{2\sigma^2}}, \]

where \( x \) is illuminance, \( \xi \) is the illuminance that the user likes most, \( \sigma \) is tolerance of the user and \( \alpha \) is maximum satisfaction of the user.

For an area \( A \) where \( N_{User} \) exist with light pattern \( X \), first the area is discretized into \( N_{Tile} \) tiles. Each of these tiles should be small enough so that illuminance of one tile can be seen as a constant. Then the user comfort of entire area \( S_A(X) \) is evaluated as the weighted average user satisfaction over all tiles and all users

\[ S_A(X) = \frac{1}{N_{User}} \sum_{j=1}^{N_{User}} \frac{\sum_{i=1}^{N_{Tile}} \beta^j_i S(x_i)}{\sum_{i=1}^{N_{Tile}} \beta^j_i}, \]

where \( x_i \) is the illuminance of tile \( i \) and \( \beta^j_i \) indicates whether tile \( i \) belongs to task area of user \( j \), which is defined in (7)

\[ \beta^j_i = \begin{cases} 1, & \text{if tile } i \text{ belongs to task area of user } j; \\ 0, & \text{otherwise.} \end{cases} \]

2.3 Constraints

It is not energy-efficient to illuminate all tiles as bright as task area. However, a large variation of illuminance may decrease visual comfort and visual performance by causing transient adaptation problems. Thus the other area should provide a well-balanced illuminance distribution in the field of view. According to [3], \( I_O \), the minimum requirement for illuminance of other area, is determined by \( \overline{IT} \), the average illuminance of task area. For an office, the relationship between \( \overline{IT} \) and \( I_S \) is listed in Table 1.

Moreover, it is important not only to provide enough light level but also to ensure that light is evenly distributed. Large variations in illuminance may lead to visual stress and
Table 1: Relationship between $I_T$ and $I_O$ [3].

<table>
<thead>
<tr>
<th>$I_T$ (lux)</th>
<th>$\geq 750$</th>
<th>$500$</th>
<th>$300$</th>
<th>$\leq 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_O$ (lux)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

eye strain [4]. The uniformity of illuminance of a certain area is quantified by minimum-to-average illuminance ratio of that area, denoted as $\theta$. The requirements for $\theta$ are different for different areas. In current European standard, $\theta$ should not be less than 0.7 over task area while for other area $\theta$ should be at least 0.5 [5].

3 Adaptive Simulated Annealing

Simulated annealing (SA) is a method that is commonly used in locating the global optimum of a given function in a large search space. It is proved that for any given finite problem, the probability that the simulated annealing algorithm terminates with a global optimal solution approaches 1 [6]. In order to make the algorithm more efficient and less sensitive to pre-defined parameters, the adaptive simulated annealing (ASA) method is applied to this problem. Similar to canonical SA, following parameters in ASA need to be specially designed according to this specific problem: temperature initialization, energy function, neighborhood search procedure, acceptance probability, annealing schedule and stop criteria. The energy function $E = Energy(W)$ is the objective function of (1). The other parameters will be introduced later. Following pseudocode presents how the adaptive simulated annealing algorithm is used to optimize the dimming vector for the LED array.

Algorithm 1 Simulated Annealing

1: for $i = 1$ to $N_{Rep}$ do 
2: Initialization() 
3: for $j = 1$ to $N_T$ do 
4: for $k = 1$ to $N_{AP}$ do 
5: for $m = 1$ to $N_{Step}$ do 
6: for $n = 1$ to $N_{LED}$ do 
7: $w_{n, New} \leftarrow NewW(w_{n, Old}, v_n)$
8: $E_{New} \leftarrow Energy(W_{New})$
9: if $E_{New} < E_{Opt}$ then 
10: $W_{Opt} \leftarrow W_{New}$
11: $E_{Opt} \leftarrow E_{New}$
12: end if (Line 9)
13: $\Delta E \leftarrow E_{New} - E_{Old}$
14: $Acc \leftarrow AccOrNot(T, \Delta E)$
15: if $Acc$ is True then 
16: $W_{Old} \leftarrow W_{New}$
17: $E_{Old} \leftarrow E_{New}$
18: $u_n + +$
19: end if (Line 15)
20: end for (Line 6)
21: end for (Line 5)
22: $V \leftarrow UpdateStepVector(V, U, N_{Step})$
23: $U \leftarrow 0$
24: end for (Line 4)
25: $Stop \leftarrow StopOrNot(E_{Opt})$
26: if $Stop$ is True then 
27: Break
28: else
29: $T \leftarrow UpdateTemperature(T)$
30: end if (Line 26)
31: end for (Line 3)
32: end for (Line 1)
33: Output $W_{Opt}$ and $E_{Opt}$

In the pseudocode, $N_{LED}$ is the number of LEDs in the array; $N_{Step}$ and $N_{AP}$ are parameters used to adjust the step vector $V$; $N_T$ is the number of loops for annealing temperature; to increase the robustness of the algorithm, it runs $N_{Rep}$ times with different initial conditions.
3.1 Neighborhood Search

Starting from the previous dimming vector $W_{Old}$, the new dimming vector $W_{New}$ is randomly generated along direction $n$

$$W_{New} = W_{Old} + ye_n$$  \hspace{1cm} (8)

where $e_n$ is the unit vector of $n$th coordinate direction; $y \in [-1, 1]$ is a random variable which is generated by [7]

$$y = \text{sgn}(z - 0.5)v_n[(1 + 1/v_n)^{2z-1} - 1],$$  \hspace{1cm} (9)

where $v_n$ is the step vector along direction $n$; $z$ is a uniformly distributed random variable, i.e.,

$$z \sim U[0, 1].$$

Since (1) is an optimization problem with several constraints, it is necessary to ensure that those constraints are satisfied by new generated dimming vector. Thus if $W_{New}$ cannot satisfy all the constraints, it will be re-generated.

3.2 Acceptance Probability

After finding the new dimming vector $W_{New}$, we need to decide whether it will be accepted or not. Here the Metropolis Criterion [8] is used to determine the acceptance probability, which is widely used in many simulated annealing based algorithms. When the temperature is $T$, the acceptance probability $P_{Accept}$ can be written as

$$P_{Accept} = \begin{cases} 
1, & \text{if } \Delta E < 0; \\
e^{-\Delta E/T}, & \text{otherwise.} 
\end{cases}$$  \hspace{1cm} (10)

This function means that our algorithm always moves downhill when it finds a way to do so, irrespective of the temperature; when the energy of new dimming vector is higher than that of old dimming vector, it is still possible to accept the new dimming vector. However, that probability decreases as the temperature cools down.

3.3 Temperature Initialization

Since the typical value of $|\Delta E|$ in this problem is usually very small, the acceptance probability calculated by (10) is either equal to or very close to 1 for high temperature $T$, which is meaningless because the algorithm retrogrades to random search under this condition. As a result, it is quite important to choose an appropriate initial temperature.

In this problem, we first run several iterations with fixed step length to calculate $|\Delta E|$, the average of $|\Delta E|$. Then the initial temperature $T_0$ can be calculated by

$$T_0 = -|\Delta E|/ \ln P_{Accept_0},$$  \hspace{1cm} (11)

where $P_{Accept_0}$ is the initial acceptance probability.
3.4 Annealing Schedule

The algorithm starts with initialized temperature $T_0$, and then as the simulation proceeds, the temperature gradually decreases at each iteration. In this way, the probability that the algorithm accepts dimming vector with higher energy becomes smaller and smaller, which means finally the algorithm will converge to the optimum solution. As demonstrated in (12), temperature for next loop is defined as multiplication of current temperature and temperature constant $c_T$, which is close to but smaller than 1.

$$T_{\text{New}} = c_T T_{\text{Old}}$$

3.5 Update Step Vector

From an optimization perspective, it is effective to maintain a one-to-one rate between accepted moves and rejected moves [9]. Too many accepted moves means the trial dimming is close to the current one, thus the energy is small compared with temperature. On the other hand, too many rejected moves means the trial dimming is far away from the current one, thus possible optimum solution might be omitted. Both of the two are somewhat wasting computational effort.

In our ASA algorithm, the trial dimming vector is generated along each direction in turn and independent with other directions. A vector $U$ is used to record the number of acceptance for each direction with current step vector. The step vector $V$ is updated every $N_{\text{Step}}$ loops. In order to maintain the one-to-one rate in direction $n$, we need to maintain the ratio $u_n/N_{\text{Step}}$ around 0.5. In this paper, we try to restrict the acceptance ratio to interval [0.4, 0.6] by

$$v_{n_{\text{New}}} = \begin{cases} \min(c_{S_1} v_{n_{\text{Old}}}, v_{\text{Max}}), & \text{if } u_n/N_{\text{Step}} > 0.6; \\ \max(c_{S_2} v_{n_{\text{Old}}}, v_{\text{Min}}), & \text{if } u_n/N_{\text{Step}} < 0.4; \\ v_{n_{\text{Old}}}, & \text{otherwise}, \end{cases}$$

where $c_{S_1} > 1$ and $c_{S_2} < 1$ are two constants used to tune the step length; $v_{\text{Max}}$ and $v_{\text{Min}}$ are the maximum and minimum step length which are used to avoid explosion.

3.6 Stop Criteria

The searching process is terminated if the algorithm cannot find better solution in past $N_e$ cooling down loops, i.e., if

$$|E_{\text{Opt}}^j - E_{\text{Opt}}^{j-N_e}| < \epsilon,$$

then stop the search. Here $\epsilon$ is a very small positive number.

3.7 Result

In this paper, the ASA is applied to three scenarios: a small (3m by 3m) office which accommodate one user and three users, a big (3m by 7m) office which contains the same three users; the height of both offices are 2.6m. Setup of the three scenarios are shown in Fig. 2a, 2b and 2c, respectively. The size of each tile is 0.1m by 0.1m.

Solving (1) with $\gamma = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ for these scenarios, the trade-off curves between increasing user comfort and reducing energy consumption for each scenario can be acquired, which are shown in Fig. 3. The values of parameters used in ASA algorithm are listed in Table 2. It is clearly seen that in the same office, the curve for three users is beneath the curve for single user, which means that it is more difficult to satisfy three users than one.
Figure 2: Setup of the three scenarios. Each of the users locates behind a desk of size 0.75m by 1.50m, with a height of 1m. The size of each task area is 0.5m by 0.5m. For every LED, $m = 25$ and $I_0 = 800$ lumen. Parameters in satisfaction functions of the three users: $\alpha = \{1,1,1\}$, $\xi = \{429, 600, 500\}$ lux, $\sigma = \{0.42, 0.42, 0.42\}$.

Moreover, for the same three users, the curve for large office is above the curve for small office, which means that for the crowded office, user’s preference of lighting condition will affect each other. This is consistent with our life experience.

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Name</th>
<th>Value</th>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{Rep}$</td>
<td>10</td>
<td>$N_T$</td>
<td>100</td>
<td>$N_{AP}$</td>
<td>20</td>
</tr>
<tr>
<td>$N_{Step}$</td>
<td>20</td>
<td>$c_T$</td>
<td>0.9</td>
<td>$c_{S_1}$</td>
<td>1.1</td>
</tr>
<tr>
<td>$v_{Max}$</td>
<td>1000</td>
<td>$c_{S_2}$</td>
<td>0.9</td>
<td>$v_{Min}$</td>
<td>0.001</td>
</tr>
<tr>
<td>$P_{Accept}$</td>
<td>0.5</td>
<td>$N_\epsilon$</td>
<td>10</td>
<td>$\epsilon$</td>
<td>$10^{-6}$</td>
</tr>
</tbody>
</table>

4 Conclusion

In this paper, we formulate a method to optimize the dimming vector of a dimmable LED array. Both increasing user comfort and reducing energy consumption are taken into consideration so that the optimized dimming vector $W_{Opt}$ is a compromise between the two aspects. A parameter $\gamma$ is used to tune the importance of the two aspects in the objective function. The other requirement such as non-task area illuminance and spacial uniformity are treated as constraints in the optimization. In order to explore the global minimum of the objective function with several constraints, a solution based on adaptive simulated annealing is proposed. As the development of LED technology, we believe that smart dimming control for a dimmable LED will be a challenging and potentially fruitful new research area.

References


Figure 3: Trade-off curve between increasing user comfort and reducing energy consumption.


Coding for known erasure values at the decoder

Ludo Tolhuizen
Philips Group Innovation, Research, Eindhoven, The Netherlands

1 Introduction

We consider the following communication situation. A sender transmits a word \( x \in Q^n \), where \( Q \) is a finite alphabet with \( q \) letters. The receiver knows a set \( E \subset Q \) of potential erasure values. Any transmitted symbol outside \( E \) is received correctly; any transmitted symbol from \( E \) is received as an erasure, indicated by \( ? \). In other words, the received vector \( x(E) \) satisfies

\[
\text{for } 1 \leq i \leq n, \quad x_i(E) = \begin{cases} x_i & \text{if } x_i \in Q \setminus E, \\ ? & \text{if } x_i \in E. \end{cases}
\]

So, for example, if \( x=(1,1,2,3) \) and \( E=\{1,3,4\} \), then \( x(E)=(?,?,2,?) \).

The receiver aims to obtain the transmitted word based on the received word and its set \( E \). The transmitter, however, does not know \( E \); it needs to send words so that for many sets \( E \), reconstruction of the transmitted word \( x \) from \( x(E) \) is feasible.

Definition 1 A code \( C \subseteq Q^n \) is called \( s \) erasure value correcting if for any set \( E \subset Q \) of size at most \( s \), a receiver with \( E \) as set of potential erased values can retrieve any transmitted word \( x \) from \( C \) from \( x(E) \). In other words, \( C \) is \( s \) erasure value correcting if

\[
\text{for all } x \in C, \; y \in C \text{ and } E \subset Q \text{ with } |E| \leq s, \quad x(E) \neq y(E) \text{ if } x \neq y.
\]

Definition 2 For integers \( n, q \) and \( s \), \( A_q(n, s) \) is the maximum size of a \( q \)-ary \( s \) erasure value correcting code.

Clearly, if \( |E| = 1 \), then the receiver knows the transmitted value for each erased symbol, and so it can always find the transmitted word. That is, \( Q^n \) is 1-erasure value correcting, and \( A_q(n, 1) = q^n \). We now give an example of a 2 erasure value correcting code, that is a special case of a more general construction.

Example 1 Let \( q \) be a prime power, and let \( Q \) be the finite field with \( q \) elements. Let \( n \leq q - 1 \), and let \( w_1, w_2, \ldots, w_n \) be distinct non-zero elements from \( Q \). For each \( K \in Q \), the code \( C(K) \) defined as

\[
C(K) = \{(x_1, \ldots, x_n) \in Q^n \mid \sum_{i=1}^n w_ix_i = K \text{ and for } 1 \leq i < j \leq n, \; x_i \neq x_j\}
\]

is two erasure value correcting.
Proof. Let $x \in C(K)$ be transmitted, let $E \subset Q$ have size at most two, and let $y = x(E)$ be received. We denote by $E$ the set of erased positions. If $E = \emptyset$, then $x = y$.

If $|E| = 1$, say $E = \{j\}$, then $x_j$ can be derived from the equation $w_jx_j = K - \sum_{i \neq j} w_ix_i = K - \sum_{i \neq j} w_iy_i$.

Now assume that $|E| \geq 2$. As the entries of $x$ are distinct, $|E| \leq |E|$. It follows that $|E| = |E| = 2$, say $E = \{i, j\}$, and each of the two elements of $E$, say $a$ and $b$, occurs in $x$ exactly once. We conclude that $x_i + x_j = a + b$. We also know that

$$w_ix_i + w_jx_j = K - \sum_{k \neq \{i, j\}} w_kx_k = K - \sum_{k \neq \{i, j\}} w_ky_k.$$  

As $w_i \neq w_j$, the above system of two equations determine $x_i$ and $x_j$. □

There are $q(q - 1) \cdots (q - n + 1)$ vectors in $Q^n$ with distinct entries. Each such vector is in exactly one of the $q$ codes $C(K)$. We conclude that there is a $K$ such that $|C(K)| \geq (q - 1)(q - 2) \cdots (q - n + 1)$.

The contents of the remainder of the paper is as follows. In Section 2, we provide some elementary properties of the function $A_q(n, s)$. In Section 3, we generalize Example 1 to arbitrary $s$ by taking proper subcodes of (shortened) Reed-Solomon codes. In Section 4, we describe a distance measure $\delta$ such that a code is $s$-erasure value correcting if and only if its minimum $\delta$-distance is at least $s + 1$. We relate the $\delta$-distance to the Hamming distance, and in this way obtain upper bounds on $A_q(n, s)$. In Section 5, we improve on the construction from Example 1 by imposing additional constraints on $w_1, \ldots, w_n$.

## 2 Elementary properties of $A_q(n,s)$

Each component of a received vector $x(E)$ can attain $(|Q| - |E| + 1)$ values, viz. an erasure, and the elements of $Q \setminus E$. As a consequence,

$$A_q(n, s) \leq (q - s + 1)^n \tag{1}$$

As we have seen before, $A_q(n, 1) = q^n$, so equality holds in (1) for $s = 1$.

Assume that $A \subset Q^n$ and $B \subset Q^m$ both are $s$ erasure value correcting. Then clearly

$$A|B = \{(a_1, \ldots, a_n, b_1, \ldots, b_m) \mid (a_1, \ldots, a_n) \in A, (b_1, \ldots, b_m) \in B\}$$

is an $s$ erasure value correcting code of length $m + n$ with cardinality $|A||B|$. As a result

$$A_q(n + m, s) \geq A_q(n, s)A_q(m, s).$$

As $A_q(n, s) \leq q^n$, it follows from Fekete’s lemma [1, Lemma 11.6] that for fixed $q$ and $s$, there exists a real number $\alpha(q, s) \leq 1$ such that

$$\lim_{n \to \infty} \frac{1}{n} \log_q(A_q(n, s)) = \alpha(q, s).$$

It could be an interesting theoretical problem to determine bounds on $\alpha(q, s)$. It follows from (1) that $\alpha(q, s) \leq \log_q(q - s + 1)$, but we believe this bound is far from sharp if $s \neq 1$. 


3 Construction of $s$ erasure value correcting codes

In this section, we generalize the code from Example 1 to general values of $s$.

**Theorem 1** Let $q$ be a prime power, and let $Q$ be the finite field with $q$ elements and with primitive element $\alpha$. Let $n \leq q - 1$, and let $C$ be the (shortened) Reed-Solomon code with zeroes $\alpha, \alpha^2, \ldots, \alpha^s$, so

$$C = \{(c_0, c_1, \ldots, c_{n-1}) \in Q^n \mid \sum_{i=0}^{n-1} c_i \alpha^{ij} = 0 \text{ for } 1 \leq j \leq s\}.$$

The code $D$ defined as

$$D = \{(c_0, c_1, \ldots, c_{n-1}) \in C \mid \text{for } 0 \leq i, j \leq n - 1, \ c_i \neq c_j \text{ if } i \neq j\}$$

is $s + 1$ erasure-value correcting.

**Proof.** Let $c \in C$ be transmitted and let $y$ be received, and let $E \subset Q$ be the set of potential erasure values of size at most $s + 1$. Moreover, we denote the set of erased positions in $y$ by $E$. We show how $c$ can be recovered from $y$ and $E$.

If $|E| \leq s$, then $c$ can be found using an erasures-only decoder for $C$.

Now assume that $|E| \geq s + 1$. As $c$ has distinct elements, $|E| \leq |E|$. As $|E| \leq s + 1$, we conclude that $|E| = |E| = s + 1$, and that $c$ contains each element of $E$ exactly once. As a consequence,

$$\sum_{i \in E} c_i = \sum_{f \in E} f.$$

So we have an additional parity check equation for the erased values in $y$, which enables us to retrieve $c$ with an erasures-only decoder for the Reed-Solomon code with zeroes $\alpha^0, \alpha, \ldots, \alpha^s$. \[\square\]

Obviously, the above theorem remains valid if we choose elements $k_1, \ldots, k_s \in Q$ and replace $C$ by $C(k_1, \ldots, k_s)$ defined as

$$C(k_1, \ldots, k_s) = \{(c_0, c_1, \ldots, c_{n-1}) \in Q^n \mid \sum_{i=0}^{n-1} c_i \alpha^{ij} = k_j \text{ for } 1 \leq j \leq s\}.$$ 

There are $q^s$ choices for $(k_1, \ldots, k_s)$, and each $x \in Q^n$ is in exactly one code $C(k_1, \ldots, k_s)$. As $Q^n$ contains $q(q-1) \ldots (q-n+1)$ with distinct entries, the following theorem holds.

**Theorem 2** For any prime power $q$ and integers $s, n$ with $s \leq n \leq q - 1$, we have that $A_q(n, s) \geq \frac{q(q-1) \ldots (q-n+1)}{q^s}$. 
4 A distance measure

In this section, we derive a distance measure that can be used to decide if a code is erasure value correcting.

Definition 3 For \( x, y \in Q^n \), we define

\[
\Delta(x, y) = \bigcup_{i:x_i \neq y_i} \{x_i, y_i\} \quad \text{and} \quad \delta(x, y) := |\Delta(x, y)|.
\]

Lemma 1 Let \( x, y \in Q^n \), and let \( \mathcal{E} \subset Q \). Then \( x(\mathcal{E}) = y(\mathcal{E}) \) if and only if \( \mathcal{E} \supseteq \Delta(x, y) \).

Proof. Let \( 1 \leq i \leq n \). If \( x_i = y_i \), then \( x_i(\mathcal{E}) = y_i(\mathcal{E}) \). If \( x_i \neq y_i \), then \( x_i(\mathcal{E}) = y_i(\mathcal{E}) \) if and only if \( x_i \) and \( y_i \) both are in \( \mathcal{E} \).

\[\square\]

Lemma 2 For \( x, y, z \in Q^n \), we have that \( \delta(x, z) \leq \delta(x, y) + \delta(y, z) \).

Proof. We will show that \( \Delta(x, z) \subseteq \Delta(x, y) \cup \Delta(y, z) \), which clearly implies the lemma. So let \( f \in \Delta(x, z) \). Let \( i \) be such that \( x_i \neq z_i \) and \( f \in \{x_i, z_i\} \). There are three cases to consider for \( y_i \).

If \( y_i \notin \{x_i, z_i\} \), then \( f \in \Delta(x, y) \) (if \( f = x_i \)) or \( f \in \Delta(y, z) \) (if \( f = z_i \)).

If \( y_i = x_i \), then there are two cases. If \( f = x_i \), then \( f = y_i \) and \( f \neq z_i \), so \( f \in \Delta(y, z) \).

If \( f \neq x_i \), then \( f = z_i \), so also in this case, \( f \in \Delta(y, z) \).

Similarly, if \( y_i = z_i \), then \( f \in \Delta(x, y) \). \( \square \)

By combining Lemma 1 and Lemma 2, we obtain the following theorem.

Theorem 3 A code is \( s \) erasure-value correcting if and only if its minimum \( \delta \) distance is at least \( s + 1 \).

The following lemma relates the \( \delta \)-distance between \( x \) and \( y \) to their Hamming distance \( d_H(x, y) \).

Lemma 3 Let \( x, y \in Q^n \). Then \( \delta(x, y) \leq q \) and \( \delta(x, y) \leq 2d_H(x, y) \).

If \( x \) has \( t \) distinct entries, then \( \delta(x, y) \geq d_H(x, y) - (n - t) \).

Proof. The two initial inequalities readily follow from the definition of \( \delta \).

In order to show the final inequality, let \( J \subset \{1, 2, \ldots, n\} \) have size \( t \) and be such that \( \{x_i \mid 1 \leq i \leq n\} = \{x_j \mid j \in J\} \). For any \( I \subset \{1, \ldots, n\} \), we have that

\[
|\{x_i \mid i \in I\}| \geq |\{x_j \mid j \in I \cap J\}| = |I \cap J| = |I| + |J| - |I \cup J| \geq |I| + |J| - n = |I| + t - n.
\]

Hence, let \( y \in Q^n \), and let \( I := \{i \mid x_i \neq y_i\} \). Then we have that

\[
\delta(x, y) = |\bigcup_{i \in I} \{x_i, y_i\}| \geq |\bigcup_{i \in I} \{x_i\}| \geq |I| + t - n = d_H(x, y) + t - n. \quad \square
\]

Lemma 3 and Theorem 3 have the following consequence.

Corollary 1 An \( s \) erasure value correcting code has minimum Hamming distance at least \( \lceil \frac{s+1}{2} \rceil \). Consequently, using the Singleton bound, \( A_q(n, s) \leq q^n - \lceil \frac{s+1}{2} \rceil = q^n - \lceil \frac{s+1}{2} \rceil \).
5 Further results for s=2

Corollary 1 implies that $A_q(n, 2) \leq q^{n-1}$. In this section we present existence results that are obtained from a closer examination of the construction from Example 1. The idea is that we need not require that all entries of the codewords be distinct if $w_1, \ldots, w_n$ satisfy additional requirements, thus possibly yielding larger codes.

**Definition 4** Let $0 \leq m \leq M \leq n$ and let $Q$ be a finite field. The vector $w = (w_1, w_2, \ldots, w_n) \in Q^n$ is called $(M, m)$-feasible if for all $I, J \subseteq \{1, 2, \ldots, n\}$ with $I \neq J$, $|I| \leq m$ and $|J| \leq M$, we have that

$$\sum_{i \in I} w_i \neq \sum_{j \in J} w_j.$$ 

The usefulness of $(M, m)$ feasible sets is explained in the following theorem, that is proved in the appendix. We use the following notation: for $x = (x_1, \ldots, x_n) \in Q^n$ and $u \in Q$, $X_u$ denotes the set $\{i \mid x_i = u\}$.

**Theorem 4** Let $Q$ be a finite field. For $1 \leq j \leq k$, let $0 \leq m_j \leq M_j \leq n$. Let $w_1, w_2, \ldots, w_n$ be $(M_j, m_j)$-feasible whenever $1 \leq j \leq k$. Let $K \subseteq Q$ and let

$$C(K) = \{(c_1, \ldots, c_n) \in Q^n \mid \sum_{i=1}^{n} w_i c_i = K\}.$$ 

For $1 \leq j \leq k$, we define

$$X_j = \{x \in Q^n \mid \forall_{\{a, b\} \subseteq Q} \max(|X_a|, |X_b|) \leq M_j \text{ and } \min(|X_a|, |X_b|) \leq m_j\},$$

$$Y_j = \{x \in Q^n \mid \forall_{\{a, b\} \subseteq Q} |X_a| + |X_b| \leq M_j + m_j\}.$$ 

Finally, we define

$$X = \bigcup_{j=1}^{k} X_j \text{ and } Y = \bigcup_{j=1}^{k} Y_j.$$ 

The code $C(K) \cap X$ is two erasure value correcting. If $Q$ has characteristic two, then $C(K) \cap Y$ is two erasure value correcting.

We now provide a different view of $(M, m)$-feasibility. Let $Q$ be a finite field with $q = p^r$ elements, where $p$ is a prime. Let $B$ be a fixed basis for $Q$ over $GF(p)$. Each element of $Q$ can be represented as a vector over $GF(p)$ of length $r$. A feasible vector $w \in Q^n$ thus is equivalent to an $r \times n$ matrix $W$ over $GF(p)$ with the following property: for all $I, J \subseteq \{1, \ldots, n\}$ with $I \neq J$, $|I| \leq m$ and $|J| \leq M$, we have that

$$\sum_{i \in I} w_{si} \neq \sum_{j \in J} w_{sj},$$

where $w_{si}$ denotes the $i$-th column of $W$. From this representation, we obtain the following results.

**Theorem 5** The columns of a binary $r \times n$ matrix $W$ correspond to an $(m, M)$-feasible vector if and only if the binary code with parity check matrix $W$ has minimum Hamming distance at least $\min(n, m + M + 1)$. 
Problem 1

C

In particular, if \( I, J \subseteq \{1, \ldots, n\} \), we have that \( \sum_{i \in I} w_i = \sum_{j \in J} w_j \) if and only if \( \sum_{k \in (I \cup J) \setminus (I \cap J)} w_k = 0 \), that is, if and only if \((I \cup J) \setminus (I \cap J)\) is the support of a non-zero codeword.

Theorem 6

Let \( W \) be an \( r \times n \) parity check matrix for \( C \). According to Theorem 5, the columns of \( W \) correspond to a vector \( (w_1, \ldots, w_n) \) in \( GF(2^r) \) that is \((M, m)\)-feasible for all pairs \((M, m)\) with \( M + m \leq d - 1 \). For \( K \subseteq Q \), let \( C(K) = \{(c_1, \ldots, c_n) \in Q^n \mid \sum_{i=1}^n w_i c_i = K\} \). According to Theorem 4, \( C(K) \cap Y \) is two erasure value correcting. As there are \( q \) choices for \( K \), at least one of the sets \( C(K) \cap Y \) has size at least \( |Y|/q \).

Corollary 2

Let \( Q = 2^r \), and let

\[
Y = \{x \in Q^n \mid \max_{\{a, b\} \subseteq Q} |X_a| + |X_b| \leq d - 1\}.
\]

If there exists a binary \([n, n - r, d]\) code, then there exists a two erasure value correcting code over \( GF(2^r) \) of size at least \( |Y|/q \).

Proof. Let \( W \) be an \( r \times n \) parity check matrix of a ternary code with minimum Hamming distance \( d \). The columns of \( W \) correspond to a vector that is \((M, m)\)-feasible for all pairs \((M, m)\) with \( M + m \leq d - 1 \).

Proof. Let \( I, J \subseteq \{1, \ldots, n\} \) be such that \( I \neq J \) and \( \sum_{i \in I} w_i = \sum_{j \in J} w_j \). We write \( I^* = I \setminus (I \cap J) \) and \( J^* = J \setminus (I \cap J) \), and \( M = \max(|I|, |J|), m = \min(|I|, |J|) \). We then have that \( \sum_{i \in I^*} w_i = \sum_{j \in J^*} w_j \), and so the vector with 1’s in the positions indexed by \( I^* \), -1’s in the positions indexed by \( J^* \) and zeroes elsewhere is a non-zero codeword, and so \( d \leq |I^*| + |J^*| \leq |I| + |J| = M + m \).

More generally, let \( W \) be a \( p \)-ary \( r \times n \) parity check matrix of a code \( C \), and let \( d^* \) be the smallest weight of a non-zero word with only entries from \([-1, 0, 1]\) in \( C \). Then \( w_1, \ldots, w_n \) is \((M, m)\)-feasible for all pairs \((M, m)\) with \( M + m \leq d^* - 1 \).

In particular, if \( C \cap \{-1, 0, 1\}^n = \{0\} \) then \( w_1, \ldots, w_n \) is \((n, n)\)-feasible.

Problem 1

Let \( p \) be a prime number, and let \( r \geq 1 \). Find the largest possible value of \( n \) such that there exist vectors \( w_1, \ldots, w_n \in (\mathbb{F}_p)^r \) such that for all \( I, J \subseteq \{1, \ldots, n\}, \) we have that

\[
\sum_{i \in I} w_i = \sum_{j \in J} w_j \text{ if and only if } I = J.
\]

In the setting of Problem 1, we have clearly have that \( 2^n \leq p^r \). For \( 1 \leq i \leq r \) and \( 0 \leq j \leq |\log_2(p)| - 1 \), let \( w_{i, j} \) be the vector for which the \( i \)-th component has entry \( 2^j \), and all other entries are zero. It is easy to check that these vectors satisfy the above condition, so we can have \( n \geq r \cdot |\log_2(p)| \).

As an example, consider Problem 1 for \( p = 7 \) and \( r = 2 \). We can readily construct 4 vectors (viz., \((0,1),(1,0),(0,2),(2,0)\)), while there can be at most \( n = 5 \) vectors. In fact, the five columns of the following matrix \( W \) satisfy the conditions of Problem 1:

\[
W = \begin{pmatrix}
1 & 4 & 2 & 0 & 1 \\
3 & 1 & 1 & 0 & 0
\end{pmatrix}
\]
We remark that Problem 1 can also be stated as finding a parity check matrix for a \( p \)-ary code correcting all error patterns consisting of symbols 0 or 1, or, equivalently, detecting all error patterns consisting of symbols from \( \{-1, 0, 1\} \). In [2], the authors introduced codes for correcting all unidirectional errors of a limited magnitude (in Problem 1: magnitude 1). In [3], optimal \textit{systematic} codes for correcting all such errors were introduced; these codes, however, are not linear.

**Acknowledgements**

I wish to thank my colleagues Stan Baggen for introducing me to this problem and for many fruitful discussions, and Ronald Rietman for providing me with the example with Problem 1.

**References**


**Appendix: proof of Theorem 4**

Let \( K, w_1, w_2, \ldots, w_n \) be in \( Q \), and let \( x \in Q^n \) and \( y \in Q^n \) be such that

\[
\sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} w_i y_i = K.
\]

Let \( a \) and \( b \) be two distinct elements in \( Q \) such that that \( x(\{a, b\}) = y(\{a, b\}) = z \). For \( u \in Q \), we define

\[
X_u = \{ i \mid x_i = u \} \quad \text{and} \quad Y_u = \{ i \mid y_i = u \}.
\]

We have that

\[
0 = \sum_{i=1}^{n} w_i x_i - \sum_{i=1}^{n} w_i y_i = \sum_{i \in X_a \cup X_b} w_i x_i - \sum_{i \in Y_a \cup Y_b} w_i y_i,
\]

where the second equality holds as \( x_i = y_i \) if \( x_i \notin \{a, b\} \) or \( y_i \notin \{a, b\} \). Now note that \( x_i = y_i = a \) if \( i \in X_a \cap Y_a \) and \( x_i = y_i = b \) if \( i \in X_b \cap Y_b \). We thus have that

\[
0 = a \sum_{i \in X_a \cap Y_b} w_i + b \sum_{i \in X_b \cap Y_a} w_i - a \sum_{i \in Y_a \cap X_b} w_i - b \sum_{i \in Y_b \cap X_a} w_i.
\]
and so

\[ 0 = (a - b) \left( \sum_{i \in X \cap Y_b} w_i - \sum_{i \in X \cap Y_a} w_i \right). \]

As \( a \neq b \), we conclude that

\[ \sum_{i \in X \cap Y_b} w_i = \sum_{i \in X \cap Y_a} w_i. \tag{2} \]

Now suppose that \( |X_a| \leq m_j, |X_b| \leq M_j \) for some \( j \leq k \). Then we clearly have that \( |X_a \cap Y_b| \leq |X_a| \leq m_j \) and \( |X_b \cap Y_a| \leq |X_b| \leq M_j \). If \( w \) is \((M_j, m_j)\)-feasible, it follows from (2) that \( X_a \cap Y_b = Y_a \cap X_b \). This implies that \( X_a \cap Y_b = Y_a \cap X_b = \emptyset \) (if \( i \in X_a \cap Y_b = Y_a \cap X_b \), then \( i \) is in \( X_a \), so \( x_i = a \), and \( i \) is in \( X_b \), so \( x_i = b \), contradicting that \( a \neq b \)). As \( x(\{a, b\}) = y(\{a, b\}) \), we infer that \( x = y \). This proves the first part of Theorem 4.

We now assume that \( Q \) has characteristic two. Then (2) is true if and only if \( \sum_{i \in I} w_i = 0 \), where \( I = (X_a \cap Y_b) \cup (X_b \cap Y_a) \). Now suppose that \( x \in Y \), so \( |X_a| + |X_b| \leq M_j + m_j \) for some \( j \). Then surely \( |I| \leq M_j + m_j \). Let \( A \) consist of the \( \min(M_j, |I|) \) smallest elements from \( I \), and let \( B := I \setminus A \). Then \( |B| \leq m_j \), and

\[ \sum_{i \in I} w_i = \sum_{i \in A} w_i + \sum_{i \in B} w_i = \sum_{i \in A} w_i - \sum_{i \in B} w_i \neq 0, \]

where the inequality follows from the facts \( w \) is \((M_j, m_j)\)-feasible, \( A \neq \emptyset \) and \( A \cap B \neq \emptyset \), whence \( A \neq B \).
Information Transmission using Illumination Systems

Anagnostis Tsiatmas\textsuperscript{1,2}, Constant P. M. J. Baggen\textsuperscript{2}, and Frans M. J. Willems\textsuperscript{1}

\textsuperscript{1}Department of Electrical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands  
\textsuperscript{2}Philips Research Eindhoven, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands  
emails: a.tsiatmas@tue.nl, stan.baggen@philips.com, f.m.j.willems@tue.nl

Abstract

In this paper, we present the Square Root (SR) channel model for the description of optical IM/DD systems. We compare the SR channel with other optical channel models and present bounds on its capacity under peak power, average power, and bandwidth constraints. Based on the SR channel model, we investigate the power efficiency and information capacity of joint illumination and communication systems.

1 Introduction

Traditionally, light-emitting-diodes (LEDs) and laser diodes (LDs) have been exploited for data transmission in optical fibers through light modulation. Fiber communication systems are widely used in the present backbone communication networks, since they can provide propagation distances in the order of hundreds of kilometres and transmission data-rates up to several Gbps. Similarly, in the late 70’s Gfeller and Bapst showed that light modulation can be used for optical wireless (OW) communications \cite{1}. OW systems using infrared (IR) LED sources gained a lot of attention in the past, since they have a number of attractive qualities, such as a license-free electromagnetic spectral range of hundreds of THz, a high spatial confinement due to the inability of IR radiation to penetrate walls, and the potential to achieve high data-rates. However, in free-space propagation contrary to fibers, losses are high and limit the applicability of OW systems in commercial products. Recently, the emergence of white high-brightness LEDs (HB-LEDs), which allow for simultaneous operation as illumination and communication sources, renewed the interest in OW communications, but with the focus on visible wavelengths this time.

Many of the optical fiber or wireless communication systems employ intensity-modulation/direct-detection (IM/DD). IM/DD is particularly popular with low-cost optical sources that do not allow for phase control of the light, such as LEDs. More precisely, at the transmitter, non-negative input signals modulate the intensity of the emitted wave, which propagates towards the receiver. There, a photodetector (PD) is used to convert linearly the optical intensity (i.e., the number of received photons) of the radiation incident on its surface into an output current signal (i.e., a number of generated electrons).

This paper is divided into two parts. In the first part, we present the Square Root (SR) channel and the derived bounds on the capacity of this channel under peak power, average power, and bandwidth constraints. The SR channel is a general channel model that can be used in IM/DD systems. In the second part, we examine the interesting case of joint illumination and communication systems. For these systems, we apply the SR channel model to evaluate their performance and efficiency.
2 The Square Root Channel

2.1 Channel Model and Capacity Results

IM/DD channels are usually modeled using Poisson processes, as a result of both the quantum nature of the light generation process in the source and the electron generation process in the PD. Therefore, only the average and not the exact number of the generated electrons over a time-interval of duration $T$ can be determined a-priori. Unless stated differently, we assume $T = 1$.

**Definition 1:** The memoryless discrete-time Poisson channel (Fig. 1(a)) has as input the r.v. $\Lambda \geq 0$, i.e., the average rate of generated photoelectrons in the receiver due to the received signal, and as output the discrete r.v. $X$, which is drawn from a Poisson distribution of parameter $\Lambda + \lambda_0$ ($X \sim \mathcal{P}(\Lambda + \lambda_0)$). The non-negative constant $\lambda_0$ is related to the ambient light and the random generation of electrons in the PD [2]. The Poisson channel law for input $\Lambda = \lambda$ and output $X = x$ is:

$$w_{\text{P}}(x|\lambda) = \exp\left(-\lambda - \lambda_0\right) \frac{(\lambda + \lambda_0)^x}{x!}, \quad x \in \mathbb{N}, \text{ and } \lambda \geq 0.$$  \hspace{1cm} (1)

In this description, the received signal is stochastic with its variance depending both on the signal level and the parameter $\lambda_0$. The dependence of the variance on the signal value complicates the study of Poisson channels, including the efficient signal-design and the derivation of decoding methods for these channels. On the other hand, Additive White Gaussian Noise (AWGN) channels have been examined extensively by the scientific community, and the results of these studies are well-understood. For this reason, in [3] we have introduced the Square Root (SR) channel for the description of the optical IM/DD systems. The SR channel is, in essence, an AWGN channel.

**Definition 2:** The memoryless discrete-time Square Root (SR) channel (Fig. 1(b)) has as input the r.v. $\Lambda \geq 0$, i.e., the average rate of generated photoelectrons in the receiver due to the received signal, and as output the r.v. $Y$, which is defined as $Y = \sqrt{\Lambda + \lambda_0} + N$. The non-negative constant $\lambda_0$ is related to the ambient light, the thermal noise in the receiver and the random generation of electrons in the PD, while the r.v. $N$ is drawn from a Gaussian distribution of mean value equal to 0 and variance equal to $1/4$ ($N \sim \mathcal{N}(0, 1/4)$). The SR channel law for input $\Lambda = \lambda$ and output $Y = y$ is:

$$w_{\text{SR}}(y|\lambda) = \sqrt{\frac{2}{\pi}} \exp\left(-2\left(y - \sqrt{\lambda + \lambda_0}\right)^2\right), \quad y \in \mathbb{R}, \text{ and } \lambda \geq 0.$$  \hspace{1cm} (2)

The SR channel is inspired by the initial empirical observation by Barlett [4] and the subsequent proof provided by Curtiss [5] that for a r.v. $X \sim \mathcal{P}(\lambda)$ and any constant $a \in \mathbb{R}$, the transformed r.v. $\sqrt{X + a}$ has asymptotically Gaussian distribution with mean value equal to $\sqrt{\lambda + a}$ and variance equal to $1/4$ as $\lambda \rightarrow \infty$.

In [3] we have shown that asymptotically with large peak or average powers, coding and decoding techniques for the SR channel can be applied to the Poisson IM/DD channel without
inducing any loss in the information transmission rate of the latter. This asymptotic behaviour is already reached for values of peak or average power in the range of 10 to 100 electrons/symbol-interval. In addition, we have seen that in IM/DD channels the PD converts the intensity of the optical wave into an output current signal. In this sense, the PD acts as a squarer and this leads to a discrepancy between the optical power of the incident wave and the electrical power of the output current. The application of the square-root transform inverts the squaring behaviour of the PD and matches the power in the optical and the electrical domains.

If we now define \( S \triangleq \sqrt{A + \lambda_0} \), we can write the SR channel equation in a simpler form: \( Y = S + N \), with \( N \sim \mathcal{N}(0, 1/4) \). This is immediately recognised as the form of an AWGN channel equation, with the only additional restriction that \( S \geq \sqrt{\lambda_0} \). Then, in the case of a peak-power constraint \( (S^2 \leq A + \lambda_0) \), the SR channel has the following lower \( (C_{\text{LB}}(A)) \) and upper bounds \( (C_{\text{UB}}(A)) \) on its capacity [3]:

\[
C_{\text{LB}}(A) = \frac{1}{2} \log_2 \left( 1 + \frac{2}{\pi e} \right),
\]

\[
C_{\text{UB}}(A) = \epsilon \log_2 (l) + \frac{1 - \epsilon}{2} \log_2 \left( \frac{\pi e}{2(1 - \epsilon)} \right) + H_2(\epsilon) - \frac{1}{2} \log_2 \left( \frac{\pi e}{2} \right),
\]

with \( l = \left( \sqrt{AT + \lambda_0 T} - \sqrt{\lambda_0 T} \right) \), \( \epsilon \) the root of the equation \( (1 - \epsilon)^2 = \frac{\pi \epsilon}{27} \), and \( H_2(p) \) the entropy in bits of the Bernoulli distribution with success probability equal to \( p \). Similarly, in the case of an average power constraint \( (E[S^2] \leq \epsilon + \lambda_0) \) and \( \lambda_0 = 0 \), we have the lower \( (C_{\text{LB}}(\epsilon)) \) and upper \( (C_{\text{UB}}(\epsilon)) \) bounds on the capacity of the SR channel [3]:

\[
C_{\text{LB}}(\epsilon) = \frac{1}{2} \log_2 (1 + \epsilon T),
\]

\[
C_{\text{UB}}(\epsilon) = \frac{\epsilon}{2} \log_2 \left( \frac{\pi e (\epsilon T + 1/4)}{2\epsilon} \right) + \frac{1 - \epsilon}{2} \log_2 \left( \frac{\pi e}{16(1 - \epsilon)} \right) + H_2(\epsilon) - \frac{1}{2} \log_2 \left( \frac{\pi e}{2} \right),
\]

where now \( \epsilon = \frac{1}{1 + \left( \frac{1}{\pi e} \right)^2} \). These bounds already for values of \( \epsilon > 10 \) electrons/symbol-interval differ by less than 0.5 bits/channel-use, with this difference becoming less than 0.2 bits/channel-use when \( \epsilon > 100 \) electrons/symbol-interval. In the next section, we will examine the significance of these values for real OW systems.

Remark 1: Using the SR channel model we can make an interesting remark regarding the power efficiency of possible modulation schemes for IM/DD systems. We define the figure-of-merit (FOM) \( R_{\text{eff}} = 1 - \left( E[S] \right)^2 / E[S^2] \). This FOM represents the portion of power contributing to communications compared to the total emitted power by the optical source. More precisely, the information content is encoded in signals that are distributed around a mean value \( E[S] \). Since the signals cannot be negative, we have that \( E[S] > 0 \). Therefore, from the average amount of power that the light source emits \( (E[S^2]) \), only the portion \( E[S^2] - (E[S])^2 \) is contributing to the information transmission, while the rest is from a communication perspective a necessary power loss in order to achieve intensity modulation (however, it could be useful for another functionality such as illumination). This observation leads to the definition of the average signal-to-noise ratio (SNR) of the SR channel as: \( \text{SNR}_{\text{SR}} = \text{Var}(S)/\sigma_N^2 \), where \( \sigma_N^2 = 1/4 \).

2.2 Signal-Independent Noise in the Square Root Channel

In OW systems, ambient light from the sun, skylight, or other artificial-light sources such as incandescent, fluorescent, and LED lamps reaches the PD. The statistical behaviour of the total number of ambient-light received photons over a specific time-interval is also described by a
Poisson distribution. Therefore, the ambient light causes an increase in both the DC component of the PD’s output current signal and the noise level in the receiver. Since the DC component of the current signal is not carrying any information, it can be easily removed. However, the noise increase is unavoidable. It can be the case, especially in IR wireless systems, that the ambient-light noise is strong and dominates over the signal-dependent noise. In these cases, a Gaussian distribution can be used to model the noise in the channel [6]. Furthermore, in low lighting conditions, i.e., when a small amount of photons are incident on the PD, the electronics in the receiver can lead to a thermal noise-limited behaviour, which is Gaussian in nature. In cases where the ambient-light or thermal noise are the dominant noise sources in OW systems, the Optical Intensity (OI) channel model is usually considered [7].

Definition 3: The memoryless discrete-time Optical Intensity (OI) channel has as input the r.v. \( Y \), and as output the r.v. \( X \), where the r.v. \( N_1 \sim \mathcal{N}(0, \lambda_0) \). The non-negative constant \( \lambda_0 \) is related to the ambient-light and the thermal noise power. The OI channel law for input \( X = \lambda \) and output \( Y = y \) is:

\[
w_{\text{OI}}(y|\lambda) = \frac{1}{\sqrt{2\pi\lambda_0}} \exp\left(-\frac{(y - \lambda)^2}{2\lambda_0}\right), \quad y \in \mathbb{R}, \text{ and } \lambda \geq 0.
\] (7)

Proposition 1: If \( \mathcal{L} \) is the set of input signals \( \Lambda \), asymptotically as \( \lambda_0/\lambda \to \infty, \forall \lambda \in \mathcal{L} \), the SNR of the SR channel is converging to the SNR of the OI channel.

Proof: We have for the SNR of the SR Channel:

\[
\text{SNR}_{\text{SR}} = \frac{\text{Var}(S)}{\sigma_N^2} = 4 \left( E[S^2] - (E[S])^2 \right) = 4 \left( E[\Lambda + \lambda_0] - (E[\sqrt{\Lambda + \lambda_0}])^2 \right), \quad (8)
\]

Since \( \lambda_0/\lambda \gg 1, \forall \lambda \in \mathcal{L} \), from Eq. (8) we obtain:

\[
\text{SNR}_{\text{SR}} = 4 \left( E[\Lambda] + \lambda_0 - \left( E[\sqrt{\lambda_0} \left( \frac{1}{2} \frac{\Lambda}{\lambda_0} - \frac{1}{8} \frac{\Lambda^2}{\lambda_0^2} + \ldots \right) \right) \right)
\]

\[
= 4 \left( E[\Lambda] + \lambda_0 - \left( \lambda_0 + \frac{1}{4} \frac{(E[\Lambda])^2}{\lambda_0} + E[\Lambda] - \frac{1}{4} \frac{E[\Lambda^2]}{\lambda_0} \right) \right) + O\left( \frac{1}{\lambda_0^2} \right)
\]

\[
\lambda_0 \to \infty, \quad \frac{E[\Lambda^2] - (E[\Lambda])^2}{\lambda_0} = \frac{\text{Var}(\Lambda)}{\lambda_0} = \text{SNR}_{\text{OI}}.
\] (9)

Remark 2: Note that this analysis applies also in systems where only small intensity variations are allowed around an average light level (i.e., small modulation indices). In this case, the \( \lambda_0 \) term corresponds to the average light power level.

Let us now examine the dependence of the SNR on the propagation distance \( d \) for the SR channel. We assume that the \( \lambda_0 \) term includes the average signal level (\( \lambda_{\text{TX}} \)), and thermal noise and ambient-light noise from other sources at the receiver (\( \lambda_{\text{RX}} \)). In this case, if we take into account that in free-space the intensity of the optical wave has an \( 1/d^2 \) dependence on \( d \), we can write: \( \text{SNR}_{\text{SR}} = 4 \text{Var}(\sqrt{\Lambda}/d^2 + \lambda_{\text{TX}}/d^2 + \lambda_{\text{RX}}) \). Therefore, there is a distance \( d^* \) such that for \( d > d^* : \lambda_{\text{RX}} > \lambda_{\text{TX}}/d^2 + \lambda/d^2, \forall \lambda \in \mathcal{L} \) (i.e., signal-independent receiver noise dominates). In this region, Eq. (9) reveals that the \( \text{SNR}_{\text{SR}} \sim 1/d^4 \). However, for \( d < d^* \) (i.e., the transmitter-induced Poisson noise is dominant) the \( \text{SNR}_{\text{SR}} \sim 1/d^2 \) (Fig. 2(a)).

2.3 Capacity Results for the Continuous Time Square Root Channel

At this stage, we investigate the capacity of the SR channel in the continuous-time domain. We focus on the estimation of lower bounds on the capacity in the case of a bandwidth constraint in addition to a peak power constraint of the form \( S^2 \leq A + \lambda_0 \).
We assume that the channel model is the one illustrated in Fig. 2(b). After the square-root transform we have an ideal low-pass filter, defined in the frequency domain as:

$$h(f) = \begin{cases} 
1 & \text{if } |f| \leq W, \\
0 & \text{otherwise},
\end{cases}$$

(10)

where $W$ is the available system bandwidth.

We start with the derivation of an approximate lower bound that, although not completely accurate, seems intuitive and well connected to our presented results. We have seen that Eq. (3) is a lower bound on the capacity in the case of a peak power constraint of the discrete-time channel. We assume that we signal with a rate of $B$ symbols/s, and we know that $B$ is roughly twice the available system bandwidth [8]. Therefore, Eq. (3) with the substitution of $T = 1/2W$ expresses the number of transmitted bits/channel-use, while at the same time we have $2W$ channel-uses/s. So, an approximate lower bound on the capacity ($C_{AE}(A;W)$) is:

$$C_{AE}(A;W) \approx W \log_2 \left( 1 + \frac{1}{\pi e} \frac{(\sqrt{A + \lambda_0} - \sqrt{\lambda_0})^2}{W} \right), \quad [\text{bps}].$$

(11)

We now define an exact lower bound following a method similar to [9].

**Proposition 2:** In the bandlimited continuous time SR channel, with available bandwidth $W$ a lower bound on the capacity ($C_{LB}(A;W)$) is:

$$C_{LB}(A;W) = W \log_2 \left( 1 + \frac{e}{\pi^3} \frac{(\sqrt{A + \lambda_0} - \sqrt{\lambda_0})^2}{W} \right), \quad [\text{bps}].$$

(12)

**Proof:** Eq. (12) is proved by following the derivation in [9] and noting that (a) $N_0/2 = \frac{1}{4}$ and (b) the maximum-entropy distribution in the case of a peak-power constraint for the input signals $b$ in the Eq. (4) in [9] is the uniform distribution on $[\sqrt{\lambda_0}, \sqrt{A + \lambda_0}]$.

3 Joint Illumination and Communication Systems

Within the domain of OW systems, we focus on visible light communications (VLC) for which the optical source of data transmission is a white HB-LED with a primary goal to provide a specific illumination level and pattern. We call this channel the optical illumination channel, since
the illumination functionality places a hard constraint on the communication functionality, not allowing the perception of illumination to change at any time. Joint illumination and communication systems are very interesting since the invested portion of energy which does not contribute to the information transmission is not wasted as it is implied by the FOM $R_{\text{eff}}$. In joint illumination and communication systems, due to the dual functionality of the source, the total emitted power is used anyhow to provide the desired illumination level.

In this section, we apply the theoretical results for the SR channel in the evaluation of the capacity of a joint illumination and communication system. We aim not only to investigate these systems but also show the general applicability of the SR channel in optical IM/DD systems.

\subsection{The Dependence of Capacity on the Bandwidth of the System}

We examine a typical joint illumination and communication system. Since the derived formulas depend on the number of generated electrons, we need to consider the power level at the receiver, the optical spectrum of the LED source, and the responsivity curve and surface area of the PD. Actually, in illumination systems, it is not the power level of the source that is of immediate importance, but its weighted version with the responsivity curve of our eyes (illuminance). Translating the illuminance level into a total number of incident photons on the PD depends on the optical spectrum of the LED source.

For a typical scenario, we assume the use of LXH8-PW50 LEDs, the optical spectra of which are presented in Fig. 3(a). For these LEDs, we have estimated that 1W of optical power is able to provide 315.7lm. Based on this conversion ratio, we can estimate the optical spectral power density (OSPD) magnitude for any illumination level. Taking into account the OSPD and assuming a PIN PD with a surface area of 7mm$^2$, we can estimate the rate of incident photons on the PD per optical wavelength. Using now the responsivity curve of the PIN PD (Fig. 3(b)), we can estimate the total electron-generation rate (Fig. 3(c)), which is used in the capacity formulas.

We also assume the use of a transimpedance amplifier (TIA) with a feedback resistance of $R_f = 100k\Omega$ in the receiver. In the examined scenario $\lambda_0 = 3.15 \times 10^{12}$ electrons/s, since the thermal noise dominates over the random electron-generation and there is no ambient light present.

For the standard requirement of an illumination level of 500lux on the desks’ surfaces in working environments, we find the electron-generation rate equal to $\lambda = 2.316 \times 10^{13}$ electrons/s. The magnitude of this value reveals that even for time-intervals of $T \sim 2.5$ns ($\sim$200MHz), which are at the limits of the HB-LEDs response capabilities, and illumination levels in the order of $\sim 0.5$lux, we still have $\lambda T \sim 10^2$ electrons. This value implies differences between the lower and upper bounds on the capacity of the channel less than 0.2 bits/channel-use and demonstrates the general applicability of our results in VLC systems.

We assume that the illumination capacity of the system (i.e., maximum output illumination level from the installed luminaires) is adequate to provide 500lux on the working surfaces of desks in offices. This assumption is justified by a minimum installation and operational cost.
3.2 The Dependence of Capacity on the Illumination Level

We can observe that the approximate lower bound \( \mathcal{C}_{AE} \) gives almost equivalent results to the actual lower bound \( \mathcal{C}_{LB} \).

3 Conclusion

We have presented the SR channel, which has the form of an AWGN channel. For the SR channel, we demonstrated bounds on its capacity under peak power, average power, and bandwidth constraints. The tightness of these bounds is adequate to allow their use in capacity estimations.
of practical systems. Using the SR channel model, we have focused on the description of joint illumination and communication systems and estimated their information capacity as a function of the available system bandwidth and of the desired illumination level. For the latter case, we have provided a modified Blahut algorithm that can account for a fixed average power constraint in the estimation of channel capacity.

References


Abstract

Signal superposition and broadcast are important features of the wireless medium. Compute-and-Forward, also known as Physical Layer Network Coding (PLNC), is a technique exploiting these features in order to improve performance of wireless networks. More precisely, it allows wireless terminals to reliably decode a linear combination of all messages, when a superposition of the messages is received through the physical medium.

In this paper, we propose a random PLNC scheme for a local interference line network in which nodes perform random access scheduling. We prove that our PLNC scheme is capacity achieving in the case of one symmetric bi-directional session with terminals on both ends of this line network model. We demonstrate that our scheme significantly outperforms any other scheme. In particular, by eligibly choosing the access rate of the random access scheduling mechanism for the network, the throughput of our PLNC scheme is at least 3.4 and 1.7 times better than traditional routing and plain network coding, respectively.

1 Introduction

Compute-and-Forward (CF), also known as Reliable Physical Layer Network Coding (PLNC) [1], is a novel Network Coding (NC) technique for wireless networks. Uncoded versions of PLNC have been considered in the literature, see e.g. [2, 3], suffering from noise accumulation along the stages of the network. By contrast, CF works with codes in such a way as to allow nodes to efficiently and reliably recover a function of the messages from multiple senders. The technique of CF shows huge improvements over traditional routing and plain NC on many network models. Among them, the line network model has been extensively studied. Most of this work has focused on deterministic, centralized scheduling [2, 4, 5]. On the other hand, for plain NC, various network models with decentralized scheduling (e.g. random access) have been studied in [6, 7, 8], where various random NC schemes have been proposed.

In this paper, we consider CF for a line network with decentralized scheduling. In particular, we consider a random access scheduling mechanism. We propose a decentralized random PLNC scheme, which is proved to be feasible and optimal in the random access scenario. The throughput of our random PLNC scheme is compared to that of traditional routing and plain NC, where a significant enhancement of performance shows up. In particular, the throughput of our random PLNC scheme is
at least 3.4 and 1.7 times better than traditional routing and plain NC, respectively. These improvements are obtained by optimizing the access rate of the random access scheduling mechanism individually for each scheme. If we use the same access rate for all schemes and let the rate approach 1, the improvement factor of the throughput of CF over the traditional routing and plain NC is even approaching infinity. This improvement is remarkably higher than the factors of 2 and 1.5, achieved by applying CF in the case of centralized scheduling.

The coding scheme that we introduce borrows some elements from the coding schemes for the line network as suggested by Pakzad et al. [9]. We will demonstrate that our coding scheme induces a network model that is within the scope of the work of Lun et al. [10]. Therefore, we will be able to resort to the maximum achievable throughput results of [10].

The paper is organized as follows. The model of a line network with random access is specified in Section 2. In Section 3 we propose our random PLNC scheme which can be embedded into the line network model with random access. Furthermore, in Section 4, the optimality of this scheme is proved by giving the capacity of the model and showing the capacity is achievable by our scheme. The comparison of our scheme to traditional routing and plain NC will also be given in Section 4 with some plots. Last but not least, we conclude our work and give our recommendations in Section 5.

2 Model

We consider a line topology modeled as a directed graph $(\mathcal{V}, \mathcal{E})$, with nodes $\mathcal{V} = \{1, 2, \ldots, M\}$ with unit distance, and edges $\mathcal{E} = \{(u, v) \mid u, v \in \mathcal{V}, |u - v| = 1\}$. Then, we build a communication model upon this topology, by considering the nodes in $\mathcal{V}$ as wireless devices. We assume time is slotted and transmitted messages are symbols from $\mathbb{F}(q) \cup \sigma$, where $\sigma$ denotes an empty transmission. Let $X_t(u)$ and $Y_t(u)$ denote the transmitted and received messages, respectively, for node $u$ in time slot $t$, and $A_t(u, v)$ the transmitted message on the directed edge $(u, v)$ in time slot $t$. The capacity of each edge is one symbol per time slot. We assume half-duplex constraints, i.e., a node cannot both transmit and receive in the same time slot. If node $u$ is not transmitting in time slot $t$ then $X_t(u) = \sigma$. For notation convenience, we sometimes use the symbol $\tau$ to denote a uniformly distributed random variable from $\mathbb{F}_q$ which is useless to the receiving node. We consider a local interference model in which the broadcast and superposition properties are now characterized as follows. For convenience, we define the nodes $\{m \mid m \leq 0 \lor m \geq M + 1\}$ as virtual nodes which are always silent.

No Broadcast: For any $u$ and $t$ it holds that

$A_t(u, u - 1) = A_t(u, u + 1) = X_t(u) = \sigma$ or

$A_t(u, u - 1) = X_t(u) \neq \sigma$ and $A_t(u, u + 1) = \tau$ or

$A_t(u, u + 1) = X_t(u) \neq \sigma$ and $A_t(u, u - 1) = \tau$. \hspace{1cm} (1)

Broadcast: For any $u$ and $t$ it holds that

$A_t(u, u - 1) = A_t(u, u + 1) = X_t(u)$. \hspace{1cm} (2)
(a) traditional routing

(b) Plain Network Coding

(c) Compute-and-Forward

Figure 1: Illustration of constraints for the three transmission modes: Communication on the thick edge implies that no useful communication is possible on thin edges.

**No Superposition:** For any \( u \) and \( t \) it holds that

\[
Y_t(u) = \begin{cases} 
A_t(u-1, u) & \text{if } A_t(u+1, u) = \sigma, \\
A_t(u+1, u) & \text{if } A_t(u-1, u) = \sigma, \\
\tau & \text{if } X_t(u-1) \neq \sigma \text{ and } X_t(u+1) \neq \sigma.
\end{cases}
\]  

(3)

**Superposition:** For any \( u \) and \( t \) it holds that

\[
Y_t(u) = A_t(u-1, u) + A_t(u+1, u),
\]

where the addition is in \( \mathbb{F}_q \), with the additional rules that \( X + \sigma = X \) for any symbol \( X \), and that \( X + \tau = \tau' \) for any \( X \) and \( \tau \), where \( \tau' \) is, like \( \tau \), a uniformly distributed random variable from \( \mathbb{F}_q \) which is useless to the receiver.

Then, we characterize traditional routing, plain NC and CF as:

- **Traditional Routing:** neither broadcast nor superposition,
- **Plain Network Coding:** broadcast, but no superposition,
- **Compute-and-Forward:** both broadcast and superposition,

which can be further specified as:

**Traditional Routing:**

If \( A_t(u, u+1) \notin \{\tau, \sigma\} \), and \( Y_t(u+1) \neq \tau \), then
\[
A_t(u-1 \pm 1, u-1) \in \{\tau, \sigma\}, A_t(u-1, u) = \{\tau, \sigma\}, \\
A_t(u+1, u+1 \pm 1) = \sigma, A_t(u+2, u+2 \pm 1) = \sigma.
\]

(6)

**Plain Network Coding:**

If \( A_t(u, u+1) \notin \{\tau, \sigma\} \), and \( Y_t(u+1) \neq \tau \), then
\[
A_t(u-2, u-1) \in \{\tau, \sigma\}, A_t(u-1, u) \in \{\tau, \sigma\}, \\
A_t(u+1, u+1 \pm 1) = \sigma, A_t(u+2, u+2 \pm 1) = \sigma.
\]

(7)

**Compute-and-Forward:**

If \( A_t(u, u+1) \notin \{\tau, \sigma\} \), and \( Y_t(u+1) \neq \tau \), then
\[
A_t(u-1, u) \in \{\tau, \sigma\}, A_t(u+1, u+1 \pm 1) = \sigma.
\]

(8)
These transmission models are illustrated in Figure 1.

Next, we specify the random access scheduling mechanism. We assume all nodes apply the plain random access approach, i.e., in each time slot each node chooses its state to be “Transmitting” or “Receiving” with probability $a$ and $1 - a$ ($a \in [0, 1]$) is a fixed constant), respectively. This choice is independent of the state in other time slots and independent of the other nodes. A node can only transmit when it is “Transmitting”, and can only receive when it is “Receiving”. For traditional routing, we assume each node has equal probability of transmission to either direction.

3 The Scheme

Our scheme is based on random linear network coding. The essence of the scheme is that we guarantee that only innovative messages are transmitted. The scheme consists of the following elements:

- We assume that there are $P$ and $Q$ messages ($P, Q \to \infty$) to be transmitted by the left and right source, respectively. These messages are denoted as $\overrightarrow{X} = \{\overrightarrow{X}(1), \overrightarrow{X}(2), \ldots, \overrightarrow{X}(P)\}$ and $\overleftarrow{X} = \{\overleftarrow{X}(1), \overleftarrow{X}(2), \ldots, \overleftarrow{X}(Q)\}$, respectively.

- We assume each node keeps three buffers of sufficiently large size, denoted by $R$, $A$ and $B$, respectively. The use of the various buffers will be explained below.

- We define the messages in the buffers of node $m$ as $Y^R_m = \{Y^R_m(i) : i = 1, 2, \ldots, N^R_m\}$, $Y^A_m = \{Y^A_m(i) : i = 1, 2, \ldots, N^A_m\}$ and $Y^B_m = \{Y^B_m(i) : i = 1, 2, \ldots, N^B_m\}$. Since all the messages and all the coefficients are chosen from $\mathbb{F}_q$ and we only do linear coding, we can express the messages as

$$Y^R_m(i) = \sum_{j=1}^{P} \alpha^R_m(i, j) \overrightarrow{X}(j) + \sum_{j=1}^{Q} \beta^R_m(i, j) \overleftarrow{X}(j), \quad (9)$$

$$Y^A_m(i) = \sum_{j=1}^{P} \alpha^A_m(i, j) \overrightarrow{X}(j) + \sum_{j=1}^{Q} \beta^A_m(i, j) \overleftarrow{X}(j), \quad (10)$$

$$Y^B_m(i) = \sum_{j=1}^{P} \alpha^B_m(i, j) \overrightarrow{X}(j) + \sum_{j=1}^{Q} \beta^B_m(i, j) \overleftarrow{X}(j). \quad (11)$$

- Next, we construct a $N^R_m \times P$ matrix $\overrightarrow{H}^R_m$, by setting $\alpha^R_m(i, j)$ as the element in its $i$th row and $j$th column, and a $N^R_m \times Q$ matrix $\overleftarrow{H}^R_m$, by setting $\beta^R_m(i, j)$ as the element in its $i$th row and $j$th column. Then, similarly, we construct $\overrightarrow{H}^A_m$ and $\overleftarrow{H}^B_m$ with $\alpha^A_m(i, j)$ and $\beta^B_m(i, j)$, respectively.

- We assume that node $m$ knows the matrices $\overrightarrow{H}^R_m$ and $\overleftarrow{H}^R_m$, i.e. a node knows which linear combination of messages is being received. This can be guaranteed if we allow the coding coefficients to be communicated without compute-and-forward.
Since each node has at most two neighbours and the topology of the network is fixed, this can be achieved at negligible overhead.

Now, the scheme operates as follows. Initially, we assume that all the buffers in all nodes are empty. In each time slot all nodes perform the following steps:

**Step 1** For each node, all receptions directly enter $R$ (for the sources nodes on both ends of the network, the original messages directly enter $R$).

**Step 2** At the beginning of each time slot, each node updates the matrices $\overset{\rightarrow}{H}_m^R$, $\overset{\leftarrow}{H}_m^R$, $\overset{\rightarrow}{H}_m^A$, and $\overset{\leftarrow}{H}_m^B$. Initialize two messages $\overset{\rightarrow}{Y} = \overset{\leftarrow}{Y} = \sigma$ ($\sigma$ is defined in Section 2).

**Step 3** If $\text{rank}(\overset{\rightarrow}{H}_m^R) \leq \text{rank}(\overset{\rightarrow}{H}_m^A)$, skip this step. Otherwise, compute a random linear combination of all the $Y_m^R(i)$, denoted as $\overset{\rightarrow}{Y} = \sum_{j=1}^{P} \alpha'(j) \overset{\rightarrow}{X}(j) + \sum_{j=1}^{Q} \beta'(j) \overset{\leftarrow}{X}(j)$. We denote vector $\overset{\rightarrow}{V} = \{\alpha'(1), \alpha'(2), \ldots, \alpha'(P)\}$. If the $\overset{\rightarrow}{V}$ is not linearly independent of all the rows in $\overset{\rightarrow}{H}_m^A$, then it discards $\overset{\rightarrow}{Y}$ and regenerates another linear combination $\overset{\rightarrow}{Y}$ until $\overset{\rightarrow}{V}$ is linearly independent of all rows in $\overset{\rightarrow}{H}_m^A$.

**Step 4** Then, each node does the same check for $\overset{\leftarrow}{H}_m^R$ and $\overset{\leftarrow}{H}_m^B$, and generates $\overset{\leftarrow}{Y}$ similarly if $\text{rank}(\overset{\leftarrow}{H}_m^R) > \text{rank}(\overset{\leftarrow}{H}_m^B)$.

**Step 5** If a node $m$ is “Transmitting” in this time slot, it broadcasts $\overset{\rightarrow}{Y} + \overset{\leftarrow}{Y}$. If this transmission is successful to the right (the transmitted message is received by its right neighbor, i.e., the right neighbor is at the state “Receiving” in this time slot), then it adds $\overset{\rightarrow}{Y}$ to buffer $A$. Symmetrically, $\overset{\leftarrow}{Y}$ is added to $B$ if this transmission is successful to the left. However, if the transmission is not successful, then nothing will be added to $A$ and $B$, and the node discards $\overset{\rightarrow}{Y}$ and $\overset{\leftarrow}{Y}$.

### 4 Performance

In this section we present the key result of the current paper, the maximum achievable throughput of CF on the line network with random access.

**Theorem 1.** For a line network with random access, the capacity for both directions is $a(1-a)$ with CF. This capacity can be achieved by using the random PLNC scheme introduced in Section 2.

**Proof. Achievability:** We first consider only the session from left to right, assuming that the right source is not transmitting anything. In this case we can interpret the operation of the scheme as follows. Innovative packets are carried through the network over a series of links from left to right. These links are unreliable in the sense that due to random access and half duplex constraints they are not always available. Observe that the model that we described above is exactly the model studied by Lun et al. in [10]. Therefore, it follows directly from [10, Theorem 2] that we can achieve a rate $a(1-a)$ for the left-to-right session. Note, that the use of compute-and-forward does not require
any generalizations of the models from [10] in which only broadcast, but not compute-and-forward, is allowed. Even though our underlying model has compute-and-forward, the abstraction described above is that of innovative packets being transmitted over a directed graph.

Next, we include the other session. Observe, that the two sessions can be analyzed independently. More precisely, if both sources are transmitting packets, the flow of innovative packets for each of the sessions can be analyzed by ignoring the other session. Once enough innovative packets are collected at the receiver it can subtract the (linear combinations of) packets from the session for which it is the source and decode the packets from the other source. Therefore, we can achieve a rate \( a(1 - a) \) for both sessions.

**Optimality:** Since the network is symmetric, we focus on the capacity along one direction only. Following the Max-Flow-Min-Cut theorem, the capacity of this network is bounded by the capacity of each edge.

We consider an individual edge. With probability \( a(1 - a) \), a transmission is made and successfully received. With probability \( a^2 \), a transmission is made but failed to be received due to the receiver is at the state of “Transmitting”. With probability \( 1 - a \), the node is “Receiving”, thus no transmission along the edge.

Hence, this channel can be considered as an erasure channel with erasure probability \( 1 - a + a^2 \), the capacity of which is \( a(1 - a) \). As a result, the throughput of our random PLNC scheme is optimal.

Next, we compare the throughput of random PLNC to that of traditional routing and plain NC presented in [6].

**Theorem 2 ([6], Theorem 6).** For a line network with random access, rates of \( \frac{a(1 - a)^2}{2} \) and \( a(1 - a)^2 \) can be achieved by using traditional routing and plain NC, respectively.

Figure 2 shows the comparison of the rate of either centralized scheduling (CS) and random access (RA) between traditional routing (TR), plain NC (PNC) and CF. Accordingly, the performance of the random PLNC scheme is labeled as “CF, RA”. As observed, our random PLNC scheme significantly outperforms the traditional routing and plain NC for the random access scenario. If we allow the probability of transmission \( a \) to be adjusted to the modes, then the rates will be maximized in traditional routing and plain NC when \( a = 1/3 \). The maximum achievable rates are 0.074 and 0.148, respectively. With random PLNC, the maximum rate 0.25 is achieved when \( a = 1/2 \). Hence, the maximum achievable rates of traditional routing and plain NC are improved by factors of 3.378 and 1.689, respectively. These factors are significantly higher than for the case of centralized scheduling, which has factors of 2 and 1.5 [4].

Figure 3 shows the ratio between the rates of random access and the rates of centralized scheduling for various transmission modes. In other words, it shows the compatibility of these transmission modes with random access. As the figure shows, CF allows the random access scheduling mechanism to utilize the network in a relatively efficient way.

Figure 4 shows the improvement of the throughput of CF over traditional routing and plain NC when fixing the value of \( a \). With deterministic centralized scheduling, capacity achieving PLNC schemes have been proposed in [2] [4] on the line network,
which have improvement factors 2 and 1.5 over traditional routing and plain NC, respectively. However, on the line network with random access, with our random PLNC scheme, improvement factors of \( \frac{2}{1-a} \) and \( \frac{1}{1-a} \) are obtained, which approach infinity when \( a \) approaches 1, as shown in the figure.

5 Conclusions and Recommendations

In this paper, we have proposed a random PLNC scheme which can be used in a line network with random access scheduling, and we have proved that this scheme is optimal in this scenario. This result not only shows the feasibility of combining CF and random access, but also indicates that CF can have even greater improvement over traditional routing and plain NC than the improvements offered under centralized deterministic scheduling. This study can be used as a basis for other studies considering CF or PLNC with random access, which might consider more general configurations with less restrictive assumptions, e.g. multiple sessions, incoherent networks, two-dimensional networks, etc. Furthermore, other decentralized scheduling schemes and MAC schemes could be considered as well, especially for some MAC schemes particularly designed for PLNC. One such study is presented in [11].
Acknowledgment

This work was supported by ERC Starting Grant 259530-ComCom and by NWO Grant 612.001.107.

References


Security of Quantum-Readout PUFs against quadrature-based challenge-estimation attacks

Boris Škorić, TU Eindhoven, b.skoric@tue.nl
Allard P. Mosk, Complex Photonic Systems, Twente University
Pepijn W.H. Pinkse, MESA+ Institute for Nanotechnology

The concept of quantum-secure readout [2] of Physical Unclonable Functions (PUFs) has recently been realized experimentally [1] in an optical PUF system. We analyze the security of this system under the strongest type of classical attack: the challenge estimation attack. The adversary performs a measurement on the challenge quantum state in order to learn as much about it as he can. Using this knowledge he then tries to reconstruct the challenge and to emulate the PUF. We consider quadrature measurements, which are the most informative practical measurements known to us. We prove that even under this attack the expected number of photons detected in the verification mechanism is approximately a factor \( S+1 \) too low; here \( S \) is the Quantum Security Parameter, defined as the number of modes in the optical system divided by the number of photons in the challenge. The photon count allows for a reliable distinction between an authentic PUF and a challenge estimation attack.

The security of the scheme derives from fundamental quantum information-theoretic principles: the number of photons in the challenge is so low compared to the dimension of the state space that measurements cannot extract enough information about the quantum state.

Figure 1: Schematic overview of the setup used in [1]. The components in the dashed box are under the verifier’s control. The first Spatial Light Modulator (SLM) shapes the wavefront to create the challenge. BS is a polarizing beam splitter. SLM2 is tuned to ‘decode’ the correct response wavefront into a parallel beam. The detector counts how many photons pass through the pinhole.

References


On the Orders of Binary Matrix Pairs

Cees J.A. Jansen
Compumatica secure networks BV
PO Box 201, 5400AE Uden, The Netherlands

Abstract

Consider a square matrix $A$ with entries from $\mathbb{F}_2$ and the matrix $A + I$. Of interest are the orders of both matrices. In particular, this paper considers the problem of finding the values of the orders such that ord$(A) = $ ord$(A + I)$. It turns out that this problem is strongly connected with the theory of jump controlled linear finite state machines. Results include non-achievable order pairs, and are applicable in areas of stream cipher design, sequences and quantum error correcting codes.

Keywords: matrix order, minimal polynomial, jump control

1 Introduction

Consider a square matrix $A$ of dimension $n \times n$ with elements from $\mathbb{F}_2$, having $C_A(x)$ and $m_A(x)$ as its characteristic and minimal polynomials respectively. The characteristic polynomial has degree $n$, the degree of the minimal polynomial is at most $n$. Next, consider adding the identity matrix to $A$, then it follows from the definition of the characteristic polynomial that $C_{A+I}(x) = C_A(x+1)$. It also follows that $m_{A+I}(x) = m_A(x+1)$, because $C(x)$ contains all roots of $m(x)$, and for every irreducible polynomial $f(x)$ over $\mathbb{F}_2$, $f(x+1)$ is also irreducible. Now the question is if $p = $ ord$(A)$, what is then ord$(A + I)$? From an application point of view an even more relevant question is for which values of $p$, we have that $p = $ ord$(A) = $ ord$(A + I)$ or equivalently $p = $ ord$(m_A(x)) = $ ord$(m_A(x+1))$. The order of a polynomial $f(x)$ over $\mathbb{F}_2$ is defined as the smallest value of $e$, such that $f(x)$ divides $x^e + 1$. Finally, note that similar matrices, i.e. matrices $A$ and $B$, such that $A = BPB^{-1}$, have identical characteristic and minimal polynomials. Therefore, without loss of generality, we can treat the problem completely by considering polynomials $f(x)$ over $\mathbb{F}_2$ and look for achievable values of $p = $ ord$(f(x)) = $ ord$(f(x+1))$.

In a paper on stream cipher design, [6], $f(x+1)$ is called the dual of $f(x)$, denoted by $f^\perp(x)$. A theory of so-called Jump-Controlled Linear Finite State Machines, that are used in stream cipher cryptography is developed in [3, 4, 5] for the binary case, [7] for the binary extension field case, and [8] for the general $\mathbb{F}_q$ case. In these papers, also the interplay between the two operators $\ast$ and $\perp$, is studied, where $\ast$ denotes the reciprocating operator, i.e. $f^\ast(x) = x^n f(x^{-1})$, $n = $ deg$(f)$, yielding many interesting results on sets of polynomials resulting from applying the two operators alternatingly.
Section 2 recapitulates some of the results from [6], and discusses ways to show whether equal order dual minimal polynomials exist. In Section 3 a list is given with values of equal order pairs. Note: in later discussions it became clear that only symmetric binary matrices are of interest. Moreover, it became clear that the focus is on matrices with corresponding dual matrices having identical minimum polynomials. In Section 4 binary symmetric $8 \times 8$ matrices are considered, with the focus on their characteristic and minimum polynomials. The final conclusions are given in Section 5.

2 Orders of polynomials and their duals

In this section a number of useful results from the theory of finite fields and polynomials over finite fields will be recapitulated. Throughout the paper only polynomials over $\mathbb{F}_2$ are considered.

2.1 The order of a polynomial

In general a polynomial $f(x)$ is composite and, hence, can be written as $f(x) = f_1^{e_1}(x)f_2^{e_2}(x) \cdots f_r^{e_r}(x)$, for some $r \geq 1$ and for all $e_i \geq 1$, where all $f_i(x)$ are distinct irreducible polynomials. The order, or equivalently, the period of a polynomial is defined as the smallest integer $p$, such that $f(x)$ divides $x^p + 1$. Clearly, $p \geq \deg(f)$.

The following results are well known, see e.g. [9].

**Proposition 1** Let $f_1(x)$ and $f_2(x)$ be two polynomials with $\gcd(f_1, f_2) = 1$, and corresponding periods $p_1$ and $p_2$. The product polynomial $f_1(x)f_2(x)$ has period $\text{lcm}(p_1, p_2)$.

**Proposition 2** Let $f(x)$ be a polynomial of period $p$ with no repeated factors. The polynomial $(f(x))^n$, $n > 1$, has period $2^v p$, with $v$ such that $2^{v-1} < n \leq 2^v$.

2.2 Results from Jump Controlled LFSMs

Let $f(x)$ be a polynomial of degree $n$ and period $p$, not having 0 and 1 as its roots. Also, let $f^*(x) = x^n f(x^{-1})$ be its reciprocal and $f^\perp(x) = f(x + 1)$ be its dual. The period of the reciprocal polynomial is also $p$, but the period of the dual may be different from $p$. Note that in general $f^{*\perp} \neq f^{\perp *}$, but $f^{*\perp *} \equiv f^{\perp *\perp}$.

In [6] it is shown that alternately applying the reciprocal and dual operators on a polynomial gives rise to a set of at most six polynomials, denoted by $S_6$-set. As the $*$ operator is period preserving but the $\perp$ operator in general is not, the number of different periods of the polynomials in an $S_6$-set is always even, and there can be 1, 2, or 3 different periods. The corresponding $S_6$-sets are denoted by $S_{6,1}$, $S_{6,2}$, and $S_{6,3}$. Sets of the type $S_{6,1}$ are also called uniform $S_6$-sets. If the period of the irreducible polynomials in a uniform $S_6$-set is maximal, so $p = 2^n - 1$, the set is called a primitive $S_6$-set. In this context, polynomials with irreducible factors of degree 1 should be avoided, as $(x + 1)^\perp = x$, which has no reciprocal.

For even degrees self-reciprocal polynomials exist, implying the existence of self-dual polynomials (and vice versa), giving rise to sets of three polynomials, denoted
by $S_3$. For irreducible polynomials of degrees $n \equiv 0 \mod 3$, $n \neq 6$, so called dual-reciprocal polynomials exist, i.e. polynomials with $f^* \equiv f^\perp$, giving rise to $S_2$-sets. The irreducible polynomial $f(x) = x^2 + x + 1$ is the only exception, in that $f \equiv f^* \equiv f^\perp$, hence forming an $S_1$-set. Note the difference for composite and irreducible polynomials: the polynomial $f(x) = (x^3 + x^2 + 1)(x^3 + x + 1)$ is self-dual and self-reciprocal, forming an $S_1$-set for degree 6. For $n = 3$ the irreducible polynomials form the first (starting from degree 2,3,...) $S_2$-set, and for $n = 4$ the first $S_3$-set occurs. The six irreducible polynomials of degree 5 are all primitive thus forming the first primitive $S_6$-set. The nine irreducible polynomials of degree 6 form an $S_6, S_2, S_3$-set. Clearly, from irreducible polynomials, infinitely many polynomials can be constructed forming $S_6, S_3, S_2, S_1$-sets.

In Chapter 5 of [6] results are given for the periods of irreducible polynomials. The following results can be obtained from the theorems in that chapter.

**Proposition 3** Any polynomial that is either in an $S_1$-set, an $S_2$-set, or a uniform $S_6$-set has a dual in the same set with the same period.

**Proposition 4** There exist polynomials in an $S_6, S_2$-set that have duals with the same period.

**Proposition 5** Any self reciprocal irreducible polynomial has a period that is less than the period of its dual. Such a polynomial forms an $S_3$-set together with its dual and the reciprocal of its dual, the latter being self dual.

**Proposition 6** Irreducible polynomials in an $S_2$-set have degree $n \equiv 0 \mod 3$, $n \neq 6$ and period dividing $2^{3n} + 2^n + 1$.

The above propositions provide us the tools to determine if there are polynomials of certain degrees and periods, that have dual polynomials with the same period.

### 2.3 Cyclotomic decomposition

Any polynomial $f(x)$ with period $e$ must be a divisor of $x^e + 1$, but not divide any $x^{e'} + 1$, with $e' < e$. Therefore, if the dual polynomial $f(x + 1)$ has the same period as $f(x)$, it must also divide $x^e + 1$. This property is trivially satisfied by self-dual polynomials, but in other cases the irreducible factors of $x^e + 1$ are required. Clearly, $x + 1$ always divides $x^e + 1$, but the other irreducible factors can be found using the results from the theory of cyclotomic polynomials, as can be found in Berlekamp’s book [1]. For $\mathbb{F}_2$ only the odd values of $e$ are of interest, as even periods are obtained from powers of irreducible polynomials. Let $Q_i(x)$ denote the $i^{th}$ cyclotomic polynomial, then the following properties hold.

\[
x^e + 1 = \prod_{d|e} Q_d(x) \tag{1}
\]

\[
\deg(Q_d) = \phi(d) \tag{2}
\]

Let $r = \text{ord}_2(d)$ denote the order of 2 modulo $d$, i.e. $r$ is the least integer such that $2^r \equiv 1 \mod d$ and $r|\phi(d)$. Then $Q_d(x)$ factors into $\phi(d)/r$ irreducible polynomials of
degree $r$. So for $p$ prime, $x^p + 1 = Q_1(x)Q_p(x) = (x + 1)(x^{p-1} + \cdots + x + 1)$ and the irreducible factors of $Q_p(x)$ are found by determining $\text{ord}_2(p)$. Note that, if $f(x)|Q_d(x)$ then also $f^*(x)|Q_d(x)$, as the reciprocal has the same period. The latter is trivially satisfied by self-reciprocal polynomials.

2.4 A result from Glaisher

Consider the expansion of $(x+1)^n$ over $\mathbb{F}_2$. The number of nonzero binomial coefficients modulo 2 depends on the number of ones in the binary representation of the integer $n$, as given by Proposition 7, which is an old result of Glaisher [2].

**Proposition 7** Let $b_2(n)$ denote the number of nonzero binomial coefficients modulo 2 in the expansion of $(x+1)^n$ over $\mathbb{F}_2$ and let $w_H(n)$ denote the number of ones in the binary representation of the integer $n$. Then $b_2(n) = 2^{w_H(n)}$.

Next, observe that $(x^n+1)/(x+1) = x^{n-1} + \cdots + x + 1$ has $n$ nonzero coefficients, and from Proposition 7 ($(x+1)^n+1)/x$ has $2^{w_H(n)}-1$ nonzero coefficients. Consequently, if $n = 2^m - 1$ for some $m$, then $(x+1)^n+1)/x$ must be equal to $(x^n+1)/(x+1)$, implying that there always exist polynomials with period $n = 2^m - 1$, such that their duals have the same period.

Another consequence of Proposition 7 is the following result.

**Proposition 8** Let $p$ be prime, not equal to $2^q - 1$ for some prime $q$ and let either $\phi(p) = \text{ord}_2(p)$ or $\phi(p) = \text{ord}_2(p)$. Then $Q_p(x) \not\equiv Q_p(x+1)$ and so there does not exist a polynomial of period $p$ whose dual has the same period.

**Proof.** As $p$ is prime, $x^p + 1 = (x+1)Q_p(x)$ and so $(x+1)^p + 1 = xQ_p(x+1)$. Also, $p \neq 2^q - 1$, so $Q_p(x+1)$ has less nonzero coefficients than $Q_p(x)$. From this, if $Q_p(x)$ is irreducible, then it cannot be also self-dual, in agreement with the $S_1$ condition mentioned before. If $Q_p(x)$ has two irreducible factors, the case of two self-reciprocals is ruled out by the same argument. Next consider the case of a reciprocal pair. Two self dual polynomials forming a reciprocal pair is impossible from [6]. A dual reciprocal pair is ruled out by the fact that the product of the pair must be $S_1$, which it is not due to the different number of nonzero coefficients. $\square$

2.5 Applying the results

How can we apply the aforementioned results to find achievable periods of pairs of polynomials $f(x)$ and $f(x+1)$ with equal periods? The following list shows a number of results.

1. The period cannot be a power of two. This is obvious: $f(x) = x^{2^k} + 1 = (x+1)^{2^k}$, so $f(x+1) = x^{2^k}$, resulting in a pair $(2^k, -)$.

2. Polynomial pairs with period $p = 2^k(2^m - 1)$ exist for $k \geq 0$ and $m \geq 2$. This is due to the factorisation of $Q_{2^m-1}(x)$, which is the product of all degree $m$ irreducible polynomials of maximum period. Raising these polynomials to some power doubles, quadruples, etc. the period, accounting for the factor $2^k$. 
3. From the previous result it follows immediately that pairs of composite polynomials exist with periods \( p = 2^k \text{lcm}((2^{m_1} - 1), (2^{m_2} - 1), \ldots) \), with \( m_i \neq m_j \) for all \( i \neq j \).

4. Self-dual polynomial pairs of degree \( m \) with period \( 2^k p_d \), with \( p_d \) a proper divisor of \( 2^m - 1 \).

5. Periods from dual-reciprocal polynomials directly apply. These polynomials can be constructed from irreducible dual-reciprocal pairs, that have non-maximal periods, which exist for degrees \( m = 3k, k > 2 \).

In the next section some values of periods are given, for which no polynomials exist such that dual pairs exist with those periods.

### 3 Impossible periods

In order to check if a period \( p \) is achievable, and not yet covered by the conditions of the previous section, the way to proceed is as follows. Only odd periods need to be considered, because if a polynomial does not satisfy the conditions, then any power of that polynomial does not satisfy the conditions either. Next, determine \( \phi(p) \) and \( \text{ord}_2(p) \) and their ratio \( r \). For prime \( p \), if \( r = 1 \), or \( r = 2 \), then Proposition 8 rules out a pair \((p, p)\). For nonprime periods \( n \), the constituting divisor periods have to be checked.

Proposition 8 actually also holds for \( Q_n(x) \), when \( n \) is not a prime, and not a divisor of \((2^{2d} + 2^d + 1)\), \( d = 3k, k > 2 \), as its factors are equal degree \( d \) irreducible polynomials.

Below is a list of odd values of period \( p \leq 103 \) for which \((p, p)\) is ruled out, as well as the reason why.

1. Proposition 8 for \( r = 1 \) rules out: 5, 9, 11, 13, 19, 25, 27, 29, 37, 53, 59, 61, 67, 81, 83, 101

2. Proposition 8 for \( r = 2 \) rules out: 17, 23, 33, 35, 41, 47, 55, 69, 71, 75, 87, 79, 97, 103

3. The Conditions \( r = 2 \) and \( d \equiv 0 \mod 3, p \) not prime, but \( p \nmid (2^{2d} + 2^d + 1) \). The following are ruled out: 39, 45, 49, 57, 77, 95, 99,

4. Period 43, with \( r = 3, d = 14 \) is ruled out, because the three irreducible polynomials are either all self-reciprocal, or one is self-reciprocal and the other two are each other’s reciprocals. The first case is ruled out the same way as for \( r = 1 \). In the second case the reciprocal pair cannot also be a dual pair, because \( d \) is not divisible by 3. If the self-reciprocal and one of the others form a dual pair, then the three irreducible polynomials must form an \( S_3 \)-set, which is ruled out by the fact that all three have the same period. Hence \((43, 43)\) is also ruled out.

5. Period 51 = 3 \cdot 17, with \( r = 4, d = 8 \). The four irreducible polynomials of degree \( 18 \equiv 0 \mod 3 \) form two reciprocal pairs from two different \( S_6,3 \)-sets\(^1\), and so \((51, 51)\) is ruled out.

---

\(^1\)Looked up in the author’s database of polynomial sets
6. Period $65 = 5 \cdot 13$, with $r = 4, d = 12$. The four irreducible polynomials of degree $12 \equiv 0 \mod 3$ are all self-reciprocal\(^1\), thus ruling out $(65,65)$.

7. Period $89$ is prime, but has $r = 8, d = 11$. The eight irreducible polynomials of degree $11$ are reciprocal pairs from four distinct $S_{6,2}$-sets\(^1\), thereby ruling out $(89,89)$.

8. Period $91 = 7 \cdot 13$, and has $r = 6, d = 12 \equiv 0 \mod 3$. The six irreducible polynomials of degree $12$ are reciprocal pairs from three distinct $S_{6,2}$-sets\(^1\), thereby ruling out $(91,91)$.

Clearly, for higher values of the period, more special cases will show up, that need closer examination.

4 Symmetric binary matrices

We investigate the situation for symmetric binary matrices with dimensions of $8 \times 8$. The restrictions on the matrices are the following, considered consecutively.

**EqOrd** The matrix and its dual have equal orders, $\text{ord}(A) = \text{ord}(A + I)$.

**IdMP** The minimum polynomials of the matrix and its dual are identical, $m_A(x) = m_A(x + 1)$.

**FinOrd** The orders of the matrix and its dual are finite, implying that the minimum polynomials do not have factors $x$ and $x + 1$.

To clarify the above restrictions, consider matrices with the following minimum polynomials: $f(x) = x^3 + x + 1$, $g(x) = x^4 + x + 1$, and $h(x) = x(x + 1)$. It can be seen that $f$ satisfies **EqOrd** and **FinOrd**, but not **IdMP**, $g$ satisfies all three restrictions, and $h$ satisfies **EqOrd** and **IdMP**, but not **FinOrd**.

The minimum polynomials of the $8 \times 8$ matrices considered must be self-dual and not have factors of degree $1$. Dimension $8$ is especially interesting, as there are precisely two $S_3$ sets and, therefore, there exist only two self-dual polynomials of degree $8$. One of these polynomials has period $255$ and the other one period $85$. Also, the three irreducible polynomials of degree $4$ form an $S_3$ set, and, hence, one polynomial is self-dual, and the product of the other two is also self-dual. Table 1 lists the characteristic and minimum polynomials of self-dual $8 \times 8$ symmetric binary matrices.

For matrices of dimensions $2 \times 2$ through $10 \times 10$ Table 2 lists all achievable equal order pairs. **Note** that the table focusses on equal order pairs, not on self-dual minimal polynomials. As we have seen above the latter condition is much more restrictive.

5 Conclusions

This paper shows that it is a relatively straightforward exercise to find achievable orders of binary symmetric matrices over $\mathbb{F}_2$, building on the mathematical foundations of the theory of jump controlled linear finite state machines. Except in a few cases where the
Table 1: Characteristic and minimum polynomials, and orders of 8 × 8 self-dual symmetric binary matrices

<table>
<thead>
<tr>
<th>$C(x)$</th>
<th>$m(x)$</th>
<th>Ord</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(x^2 + x + 1)^4$</td>
<td>$C(x)$</td>
<td>12</td>
</tr>
<tr>
<td>”</td>
<td>$(x^2 + x + 1)^3$</td>
<td>12</td>
</tr>
<tr>
<td>”</td>
<td>$(x^2 + x + 1)^2$</td>
<td>6</td>
</tr>
<tr>
<td>”</td>
<td>$(x^2 + x + 1)$</td>
<td>3</td>
</tr>
<tr>
<td>$(x^2 + x + 1)^2(x^4 + x + 1)$</td>
<td>$C(x)$</td>
<td>30</td>
</tr>
<tr>
<td>$(x^2 + x + 1)(x^3 + x + 1)(x^3 + x^2 + 1)$</td>
<td>$C(x)$</td>
<td>21</td>
</tr>
<tr>
<td>$(x^2 + x + 1)(x^6 + x^5 + x^3 + x^2 + 1)$</td>
<td>$C(x)$</td>
<td>63</td>
</tr>
<tr>
<td>$(x^4 + x + 1)^2$</td>
<td>$C(x)$</td>
<td>30</td>
</tr>
<tr>
<td>$(x^4 + x^3 + 1)(x^4 + x^3 + x^2 + x + 1)$</td>
<td>$C(x)$</td>
<td>15</td>
</tr>
<tr>
<td>$(x^8 + x^6 + x^5 + x^4 + x^3 + x + 1)$</td>
<td>$C(x)$</td>
<td>15</td>
</tr>
<tr>
<td>$(x^8 + x^6 + x^5 + x^3 + 1)$</td>
<td>$C(x)$</td>
<td>85</td>
</tr>
</tbody>
</table>

Author’s database of irreducible polynomial sets was consulted, the results given in this paper have all been obtained without performing lengthy calculations, and without the use of a computer.

In an internal report the author has described constructions based on the direct matrix product, permutation matrices, and symmetric matrices over the field $\mathbb{F}_4$.

Acknowledgement The author wishes to thank Prof. Matthew Parker of UiB, Norway, for bringing the topic of this paper to his attention.

References


<table>
<thead>
<tr>
<th>Dimension</th>
<th>Deg $C_A$</th>
<th>Order</th>
<th>$C_A(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 × 2</td>
<td>2</td>
<td>3</td>
<td>$x^2 + x + 1$</td>
</tr>
<tr>
<td>3 × 3</td>
<td>3</td>
<td>7</td>
<td>$i_3$</td>
</tr>
<tr>
<td>4 × 4</td>
<td>2 + 2</td>
<td>3,6</td>
<td>$(x^2 + x + 1)^2$</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>15</td>
<td>$x^4 + x + 1$</td>
</tr>
<tr>
<td>5 × 5</td>
<td>2 + 3</td>
<td>21</td>
<td>$(x^2 + x + 1)(i_3)$</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>31</td>
<td>$i_5$</td>
</tr>
<tr>
<td>6 × 6</td>
<td>2 + 2 + 2</td>
<td>3,6,12</td>
<td>$(x^2 + x + 1)^3$</td>
</tr>
<tr>
<td></td>
<td>2 + 4</td>
<td>15</td>
<td>$(i_3)(i_3)$</td>
</tr>
<tr>
<td></td>
<td>3 + 3</td>
<td>7,14</td>
<td>one of 3 $p_6$</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>63</td>
<td>$i_7$</td>
</tr>
<tr>
<td>7 × 7</td>
<td>2 + 2 + 3</td>
<td>21,42</td>
<td>$(x^2 + x + 1)^2(i_3)$</td>
</tr>
<tr>
<td></td>
<td>2 + 5</td>
<td>93</td>
<td>$(x^2 + x + 1)(i_5)$</td>
</tr>
<tr>
<td></td>
<td>3 + 4</td>
<td>105</td>
<td>$(i_3)(x^2 + x + 1)$</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>127</td>
<td>$i_7$</td>
</tr>
<tr>
<td>8 × 8</td>
<td>2 + 2 + 2 + 2</td>
<td>3,6,12</td>
<td>$(x^2 + x + 1)^4$</td>
</tr>
<tr>
<td></td>
<td>2 + 2 + 4</td>
<td>15,30</td>
<td>$(x^2 + x + 1)^2(i_4)$</td>
</tr>
<tr>
<td></td>
<td>2 + 3 + 3</td>
<td>21,42</td>
<td>$(x^2 + x + 1)(i_3)(i_3)$</td>
</tr>
<tr>
<td></td>
<td>2 + 6</td>
<td>63</td>
<td>$(i_3)(i_5)^3$</td>
</tr>
<tr>
<td></td>
<td>3 + 5</td>
<td>217</td>
<td>one of 9 $p_8$</td>
</tr>
<tr>
<td></td>
<td>4 + 4</td>
<td>15,30</td>
<td>$(x^4 + x^3 + 1)(x^4 + x^3 + x^2 + x + 1)$</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>85, 255</td>
<td>$(x^8 + x^6 + x^5 + x^4 + x^3 + x + 1)$, one of 9 $p_8$</td>
</tr>
<tr>
<td>9 × 9</td>
<td>2 + 2 + 2 + 3</td>
<td>21,42,84</td>
<td>$(x^2 + x + 1)^3(i_3)$</td>
</tr>
<tr>
<td></td>
<td>2 + 2 + 5</td>
<td>93,186</td>
<td>$(x^2 + x + 1)^2(i_5)$</td>
</tr>
<tr>
<td></td>
<td>2 + 3 + 4</td>
<td>105</td>
<td>$(x^2 + x + 1)(i_3)(i_4)$</td>
</tr>
<tr>
<td></td>
<td>2 + 7</td>
<td>381</td>
<td>$(i_3)(i_7)$</td>
</tr>
<tr>
<td></td>
<td>3 + 3 + 3</td>
<td>7,14,28</td>
<td>one of 9 $p_9$</td>
</tr>
<tr>
<td></td>
<td>3 + 6</td>
<td>63</td>
<td>$(i_3)(i_6)$ of per 9)</td>
</tr>
<tr>
<td></td>
<td>4 + 5</td>
<td>465</td>
<td>$(x^4 + x + 1)(i_5)$</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>73, 511</td>
<td>$(x^9 + x + 1)$, $(x^9 + x^8 + 1)$, and $p_9$</td>
</tr>
<tr>
<td>10 × 10</td>
<td>2 + 2 + 2 + 2 + 2</td>
<td>3,6,12,24</td>
<td>$(x^2 + x + 1)^5$</td>
</tr>
<tr>
<td></td>
<td>2 + 2 + 2 + 4</td>
<td>15,30,60</td>
<td>$(x^2 + x + 1)^3(i_4)$</td>
</tr>
<tr>
<td></td>
<td>2 + 2 + 3 + 3</td>
<td>21,42</td>
<td>$(x^2 + x + 1)^2(i_3)(i_3)$</td>
</tr>
<tr>
<td></td>
<td>2 + 2 + 6</td>
<td>63,126</td>
<td>$(x^2 + x + 1)^2(one of 3 p_6)$</td>
</tr>
<tr>
<td></td>
<td>2 + 3 + 5</td>
<td>651</td>
<td>$(x^2 + x + 1)(i_3)(i_5)$</td>
</tr>
<tr>
<td></td>
<td>2 + 4 + 4</td>
<td>15,30</td>
<td>$(x^2 + x + 1)(x^4 + x + 1)(i_4)$, and</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$(x^2 + x + 1)(x^4 + x^3 + 1)(x^4 + x^3 + x^2 + x + 1)$</td>
</tr>
<tr>
<td></td>
<td>2 + 8</td>
<td>255</td>
<td>$(i_3)(i_5)(x^4 + x + 1)$</td>
</tr>
<tr>
<td></td>
<td>3 + 3 + 4</td>
<td>105,210</td>
<td>$(i_3)(i_7)$</td>
</tr>
<tr>
<td></td>
<td>3 + 7</td>
<td>889</td>
<td>$(i_4)(p_6)$</td>
</tr>
<tr>
<td></td>
<td>4 + 6</td>
<td>315</td>
<td>$(i_5)$</td>
</tr>
<tr>
<td></td>
<td>5 + 5</td>
<td>31,62</td>
<td>one of 7 $i_{10}$ with per 341, $p_{10}$</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>341, 1023</td>
<td>one of 7 $i_{10}$ with per 341, $p_{10}$</td>
</tr>
</tbody>
</table>

Table 2: Achievable equal order pairs of binary matrices. $i_m$ denotes any irreducible polynomial of degree $m$, $p_m$ denotes any primitive polynomial, that has a dual, which is also primitive.


All codes are proper for erasure decoding

Ludo Tolhuizen, Philips Group Innovation, Research, Eindhoven, The Netherlands

In 1976, Leung and Hellman [1] showed that the misdetection probability of a code used for error detection on the binary symmetric channel need not be monotonic in the channel’s error probability. Codes which do enjoy this monotonicity property are called proper. There has been quite some research interest in finding conditions for codes being proper and in constructing families of proper codes, see for example [2] and [3] and the references therein.

In this paper, we consider the $q$-ary erasure channel with erasure probability $\epsilon \in (0, 1)$. That is, any transmitted symbol is changed to an erasure with probability $\epsilon$, and stays the same with probability $1 - \epsilon$. A decoder has as input a codeword in which some symbols are erased. We call a decoder $D$ monotonic if it has the following (very natural) property: for any erasure pattern $E$ that $D$ decodes correctly, it correctly decodes any subset of $E$ as well. We will show that with a monotonic decoder, any code is proper for erasure decoding, that is, the probability of not retrieving the transmitted codeword is monotonically non-decreasing in $\epsilon$.

To this end, we fix a code $C$ and a monotonic decoder $D$. For each $c \in C$, we denote by $A_{c,w}$ the number of patterns $E$ of $w$ erasures for which $c$ is not retrieved correctly if the positions of $c$ indexed by $E$ are erased. The probability that $c$ is not retrieved correctly, given that $c$ has been transmitted, thus equals

$$A_c(\epsilon) = \sum_{w=0}^{n} A_{c,w} \epsilon^w (1 - \epsilon)^{n-w}.$$ 

By straightforward differentiation, we find that for each $\epsilon \in (0, 1)$,

$$A_c'(\epsilon) = \sum_{w=0}^{n} A_{c,w} (we^{w-1}(1 - \epsilon)^{n-w} - (n - w)\epsilon^w(1 - \epsilon)^{n-w-1}) =$$

$$= \sum_{w=0}^{n-1} ((w+1)A_{c,w+1} - (n-w)A_{c,w}) \epsilon^w (1 - \epsilon)^{n-w-1}.$$ 

Consequently, $A_c$ is non-decreasing in $\epsilon$ if the following condition is met:

$$\text{for } 0 \leq w \leq n - 1, \text{ we have that } (w+1)A_{c,w+1} \geq (n-w)A_{c,w}. \quad (1)$$ 

The following lemma will be applied to show that Condition 1 is met for monotonic decoders.

**Lemma** Let $A$ be a collection of subsets of a set $U$ of size $n$ with the property that for any $A \in A$, every set $B$ with $A \subseteq B \subseteq U$ is also in $A$. Let $A_w$ denote the number of sets of size $w$ in $A$. Then for $0 \leq w \leq n - 1$, we have that $(w+1)A_{w+1} \geq (n-w)A_w$. 

Proof Let $0 \leq w \leq n - 1$. Let $S$ be the set defined as

$$S = \{(X, Y) \in \mathcal{A} \times \mathcal{A} \mid |X| = w, |Y| = w + 1 \text{ and } X \subset Y\}.$$  

For each $X \in \mathcal{A}$ of size $w$, all $(n - w)$ sets of size $w + 1$ containing $X$ are in $\mathcal{A}$. Therefore, $|S| = (n - w)A_w$.

Conversely, each $Y \in \mathcal{A}$ of size $w + 1$ has $w + 1$ subsets of size $w$, so it has at most $w + 1$ subsets of size $w$ that are in $\mathcal{A}$. As a consequence, $|S| \leq A_{w+1}(w + 1)$.

We now apply Lemma 1 with $\mathcal{A}$ equal to the set of all erasure patterns that $D$ does not decode correctly if $c$ is transmitted. As we assume that $D$ correctly decodes any subset of a correctly decoded erasure pattern, the condition of Lemma 1 is satisfied, and hence Condition 1 is met. Consequently, the probability that $D$ does not retrieve $c$, given that $c$ has been transmitted, is monotonically non-decreasing in $\epsilon$.

The overall probability $A(\epsilon)$ of not decoding correctly satisfies

$$A(\epsilon) = \sum_{c \in C} p(c) A_c(\epsilon),$$

where $p(c)$ is the a priori probability that $c \in C$ is transmitted. Our result above implies that $A$ is monotonically increasing, or is constant. In the latter case, for each $c \in C$ for which $p(c) \neq 0$, we have equality in (1) for all $w$. That is, for each $c \in C$ with $p(c) \neq 0$, either $A_{c,w} = \binom{n}{w}$ for all $w$ (if $A_{c,0} = 1$), or $A_{c,w} = 0$ for all $w$. In the former case, $c$ is never retrieved correctly; in the latter case, $c$ is always retrieved correctly, even if only erasures are received. The latter case thus can occur for at most one word in $C$ with non-zero a priori probability.

We conclude that $A$ is constant if and only if there is at most one $c \in C$ with $p(c) \neq 0$ that is always decoded correctly, while all other codewords with non-zero a priori probability are never decoded correctly (even when received without erasures). In particular, $A$ is monotonically increasing whenever $C$ has at least two words with non-zero transmission probability, $D$ is monotonic and $D$ decodes any received word without erasures to itself.

References


Indoor Propagation and Radiation Measurements for Wireless Communication Systems

Vladimir Lazov, Sofie Pollin, Guy A. E. Vandenbosch
ESAT-TELEMIC, KU Leuven Kasteelpark Arenberg 10, 3001 Leuven, Belgium
Vladimir.lazov@esat.kuleuven.be, Sofie.Pollin@esat.kuleuven.be, guy.vandenbosch@esat.kuleuven.be

Abstract

This paper contains measured data and a review of empirical models for 2.4GHz radio propagation for indoor scenarios. The classic empirical models are briefly described and the focus is placed on their application in scenarios of very dense indoor networks. Indoor propagation is one of the most complicated propagation topics based on the specific type of the building structure, used materials and obstacles. The purpose of this study is to characterize the indoor channel for IEEE 802.11g/n wireless local area networks at 2.4GHz. Today the network planning and optimization are mainly done in 2D. When moving to small cells, with the dimension of the room or even around the wireless body, 3D propagation and indoor wireless planning becomes an issue. Moreover, elevation differences between transmitter and receiver, and the impact on networking performance should be understood. A measurement based analysis of 3D indoor radio propagation is presented in this work. The results of this measurement campaigns will be used for evaluating the impact of 3D topology on indoor wireless communication performance.

Index Terms; indoor propagation; wireless communication; measurements; 3D’s network planning.

1 Introduction

We live in a connected world. It is expected that by 2025, integrated components will be approaching molecular limits and may cover complete walls [1]. For the design and maintenance of indoor wireless services the knowledge of the signal propagation in various environments is demanding. Moreover, before implementing designs and planning of wireless communication systems, accurate propagation characteristics of the environment should be well known. When cell ranges decrease from over 100m towards approximately 10m, the 3D impact of buildings and height differences of transmitter and receiver (elevation angles) become more and more relevant. Indoor propagation is one of the most
complicated propagation topics based on the specific type of the building structure; number of obstacles presented and used materials. Furthermore, objects that surround the transmitters and the receivers severely affect the propagations characteristics of any radio channel. A channel model is useful to determine the mechanism by which the indoor propagation occurs and in turns gives sufficient information to develop good communication system [4]. Models proposed to characterize the indoor propagation can be classified in two major classes: statistical (empirical) and site-specific models [1]-[3]. Both of them take different factors into consideration when computing the path loss. Deterministic models are out of the scope of this research, as we want models that can be evaluated quickly for realistic dense indoor network simulations.

The objective of this work is the performance evaluation of dense indoor networks with small cells, and the impact of 3D topology. To achieve this, a gradual approach is taken, and first the impact of 3D relative position of transmitter and receiver in an indoor setting is evaluated. The paper describes the indoor propagation measurement campaigns for 2.4GHz done in a university building, proving the theory that the elevation angle between the transmitter and the receiver has an importance in modeling and it has to be taken into account in the indoor propagation models.

Based on our measurement results derived set of propagation models is compared, and conclusions on the impact of 3D topology are discussed.

2 Path Loss Modeling State of the Art

Firstly, empirical models are the simplest approach to the prediction of the field strength in indoor environments, and allow determining the path loss as function of the distance, frequency or other conditions. The most widely used approach to compute the propagation loss between any point from the transmitter (Tx) and the Receiver (Rx) uses a simple distance-dependent formula. The path loss coefficient is often used to tune the baseline equation towards the environment, and the path loss that is relevant for a given environment is often determined empirically. Other factors, considering the impact of furniture, walls or movement, have also been proposed. To compute the path loss, one needs hence locations of the transmitter and receiver (distance) and the number of obstacles intersecting the direct ray between the transmitter and the receiver. The path loss expected from these models can be computed very fast and only simple input is needed [5],[8] Eq. (1) expresses the described distance dependent model, where $PL(d)$ is the average path loss value in dB at a transmitter-receiver separation, $d$ is the distance between the transmitter and the receiver, $n$ is the path loss exponent, $d_0$ is the reference distance of 1m. $t$ represents $\ldots$

$$PL(d) = 10 \cdot n \cdot \log_{10}(\frac{d}{d_0}) + const$$

(1)

The path loss exponent $n$ can be determine empirically for the specific environment, or the relevant value can be taken from Table 1. Table I shows different values for the path loss exponent for various environments that are well known and they are found in the literature. Values for the path loss exponent calculated from our data raw measurements are in the
range of 1.6-1.8 for indoor environment in the case of Line of Sight (LOS) between the transmitter and the receiver and with no presence of human being. On the other hand, results show increased values for the exponents in the presence of humans.

**TABLE I**

<table>
<thead>
<tr>
<th>Environment</th>
<th>Path Loss exponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free space</td>
<td>2</td>
</tr>
<tr>
<td>Urban Area</td>
<td>2.7 to 3.5</td>
</tr>
<tr>
<td>Suburban Area</td>
<td>3 to 5</td>
</tr>
<tr>
<td>Indoor (LOS)</td>
<td>1.6 to 1.8</td>
</tr>
</tbody>
</table>

The drawback of these empirical models is that the site-specific accuracy is poor especially for complicated building structures where supplementary measurement is necessary. For instance, waveguiding effects in corridors (for indoor environment scenarios) are an issue. Moreover, there are limitation in providing parameters such as delay spread and also angle of arrival (AoA). In this work, we are not interested in very detailed information such as AoA or delay spread. Nevertheless, we want to study if distance, frequency, and number of walls/floors are sufficient parameters to model dense indoor 3D wireless networks.

Another parameter that may make a change when deciding a suitable model for indoor environment is the importance of the elevation angle; which represents that angle between the transmitter and the receiver. Our aim is to show the effect of RSSI values when taking into consideration the correction factor; this would be a step forward in the process of developing or proposing a right propagation model for the indoor environment.

We determine the gain, $G$, of the receiver as follows:

$$ G = G_{\text{max}} + F(\theta), $$

(2)

where $G_{\text{max}}$ is the maximal gain [dBi] in the horizontal plane, depends on the model of the access point and the value can be read from the datasheet given by the manufacturer; $F$ depends on the elevation angle [9] and moreover is presented with the following equation, applying to thin wire dipole antennas, where

$$ F(\theta) = 10 \log_{10} \left( \frac{\cos(kL \sin(\theta) - \cos(kL))}{\cos(\theta)(1 - \cos(kL))} \right)^2, $$

(3)

$\theta$ is the angle between the transmitter and the receiver, $k = \frac{2\pi}{\lambda}$ and $2L$ is the length of the antenna. In the next section, we present in details our evaluation setup and the results where we compare the importance of elevation angle in dense 3D networks.
3 Experimental Setup

One scenario with three different positions of the transmitter and the receivers is considered in our measurements for evaluating the impact of the elevation angle. The scenario is described as follows:

- ESAT building TELEMIC lab, closed corridor is used for signal strength measurements. The corridor is closed on both sides with walls.

![Figure 1: Measurement Locations](image1)

- Signal measurements are taken with a resolution of 1m along the 13m length corridor. The transmitter is positioned at 0.8 meters above the ground, whereas the receiver was at either at 0.8m or 1.5m below or above the transmitter respectively.

![Figure 2: Schematic view of the measurement setup. The orientation of the AP and the laptop was the default one, receiver facing the AP.](image2)
In WLAN systems, one of the commonly used antennas is omnidirectional one. An omnidirectional antenna is an antenna that has a non-directional pattern (circular pattern) in a given plane with a directional pattern in any orthogonal plane. Examples of omnidirectional antennas are dipoles and collinear antennas. Namely, main importance of every antenna is the radiation pattern or antenna pattern. Antenna's pattern describes how the antenna radiates energy out into space (or how it receives energy). It is important to state that an antenna radiates energy in all directions, at least to some extent, so the antenna pattern is actually three-dimensional. Namely, in the AP used for our measurements dipole antenna is mounted so that it is vertically oriented with respect to the floor or ground [6]. This results in the maximum amount of energy radiating out into the intended coverage area. An omnidirectional antenna, which is used in our setup, is an antenna that has a non-directional pattern (circular pattern) in a given plane with a directional pattern in any orthogonal plane.

For this purpose an off-the-shelf measurement setup was designed as shown in Figure 2; The setup consisting of an Access Point (AP), type CISCO [6] which operates at 2.4GHz. A laptop with wireless client adapter/sniffer type AirPcap [7] is used to measure the signal strength. The AP uses an integrated antenna with gain 4.0 dBi, horizontal beamwidth 360°. Three scenarios are considered for the measurements, chosen for a better understanding of the 3D configuration and its impact on the network performance. Measurements are done in three different periods of the day, in the morning, afternoon and in the evening. The reason for doing this is to see if there are any differences in the parameters obtained from the results.

4 Results

For each of the measurement sets, we fit the distance dependent pathloss model, and compute the path loss exponent. Also, we fit the angle-corrected pathloss model. To compare both, we rely on the RMSE (root mean squared error)

\[
RMSE = \sqrt{\frac{\sum (f(x_i) - y_i)^2}{n}}.
\] (4)

Below, we detail the steps taken in our analysis:

1. Signal strength is measured by means of RSSI values given by our sniffer that listens to the 802.11g beacons of the AP. We average the RSSI of more than 200 beacons for each location and measurement time. The measurement points are given in Fig. 3.

2. Similarly, we compute the angle-corrected RSSI for each measurement location, by accounting for a different antenna gain for each location, and plot those results in the Figures 4 until 6. Namely, the values obtained with equation (2) will correct the received RSSI values from the measurements for all the cases.

3. On both data sets, we plot the log-distance path loss model using the least squares method, and compute the path loss exponent and RMSE.
Figure 3. Path Loss Model 3 positions (left); Curve fitting and extracting the path loss exponents for closed corridor measurements: three different heights for the Rx with and with no elevation angle correction (right).

Figure 4. Path Loss Model with corresponding path loss exponents in the afternoon (left) and evening (right) with and with no angle correction.
Figure 5. Comparing the results from the three periods of the day for the path loss model with elevation angle correction

TABLE II:
RMSE values for Path Loss Model with and with no correction.

<table>
<thead>
<tr>
<th>Period of the day</th>
<th>Log-distance Model (no correction)</th>
<th>Angle-dependent (correction)</th>
<th>Difference in percent with/without correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8m(7-8pm)</td>
<td>2.66</td>
<td>2.58</td>
<td>3.001%</td>
</tr>
<tr>
<td>1.5m(7-8pm)</td>
<td>2.85</td>
<td>1.76</td>
<td>38.24%</td>
</tr>
<tr>
<td>0.8m(1-3pm)</td>
<td>4.28</td>
<td>4.007</td>
<td>6.37%</td>
</tr>
<tr>
<td>1.5m(1-3pm)</td>
<td>3.53</td>
<td>2.85</td>
<td>19.26%</td>
</tr>
<tr>
<td>0.8m(9-10am)</td>
<td>2.26</td>
<td>2.007</td>
<td>11.19%</td>
</tr>
<tr>
<td>1.5m(9-10am)</td>
<td>4.32</td>
<td>3.74</td>
<td>13.42%</td>
</tr>
</tbody>
</table>

TABLE III:
Summarized values for the path loss exponents and RMSE values for all the points combined, obtained from the measurement data for the three periods of the day

<table>
<thead>
<tr>
<th>Period of the day</th>
<th>PL Exponent: Log-distance Model (no correction)</th>
<th>PL Exponent: Angle-elevation (correction)</th>
<th>RMSE values: No correction</th>
<th>RMSE values: With correction</th>
<th>Difference in percent with/without correction: RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morning</td>
<td>1.683</td>
<td>1.349</td>
<td>4.75</td>
<td>4.29</td>
<td>9.68%</td>
</tr>
<tr>
<td>Afternoon</td>
<td>1.404</td>
<td>1.092</td>
<td>4.57</td>
<td>4.43</td>
<td>3.06%</td>
</tr>
<tr>
<td>Evening</td>
<td>1.793</td>
<td>1.481</td>
<td>3.55</td>
<td>2.83</td>
<td>20.28%</td>
</tr>
</tbody>
</table>
Comparing the results provided in the Tables II and III and in figures, will help us proposing a suitable path loss model for an indoor communication. It is clearly that we have improvements in the path loss model including the angle correction factor. Path loss exponent is a clear indication that shows how rapidly the path loss increases with increasing the transmitter-receiver separation. Namely, its values for all the cases have smaller values with the correction included in the calculations. Moreover, there is a difference and improvement in the RMSE values and path loss exponents for all the periods of the day and positions of the transmitter and the receiver as shown in the Table III. Comparison for the path loss model with the elevation angle correction included for different periods of the day is presented in figure 5. Results gained from the measurements with the statistical approach including the angle correction factor indicate that evaluated path loss model with the elevation angle correction for a communication between the transmitter and the receiver is suitable for specific indoor scenarios.

5 Conclusion

In this study indoor propagation models are investigated for a closed corridor scenarios. The objective of this work is the performance evaluation of dense indoor networks with small cells, and the impact of 3D topology The Log-distance Path loss Model (Empirical Model) is evaluated in details and it is studied if the propagation estimates should account for the angle between the relative position of the transmitter and receiver. Line of sight communication between the transmitter and the receiver is always present in all measurement points.

Firstly, the PL exponents are calculated and extracted with line fitting curve. Values vary from 1.404-1.793 and they are summarized in Table III. Clear indication of the importance for including the elevation angle as correction factor in the modeling is the obtained smaller values for the path loss exponents ranging from 1.09 to 1.48 compared to those without correction. Secondly, the RMSE values are smaller for all the cases. Namely, the improvements and the results presented in the Tables II and III, the elevation angle between the transmitter and the receiver has an importance in modeling and it has to be taken into account in the indoor propagation models.

Although the log-normal propagation model is a popular model for wireless networks, there are environments in which the log-distance propagation model is not the best one selected. In that case, a technique and further research will be held using the same scenario as previously described in order to improve and propose best model between the transmitter and receiver in indoor environment communication.

6 References


Designated Attribute Proofs with the Camenisch–Lysyanskaya Signature

Kostas Papagiannopoulos, Gergely Alpár, and Wouter Lueks
Institute for Computing and Information Sciences (iCIS), Radboud University Nijmegen, The Netherlands.

Abstract. An attribute-based credential (ABC), an authentic personal electronic piece of information to perform transactions, offers a secure and privacy-preserving method to authorize user transactions. One of the most important techniques to realize ABCs is the Camenisch–Lysyanskaya signature. In this paper we extend its underlying cryptographic algorithms with the designation property. As a result, the showing protocol between a user and a verifier does not require an additional secure channel to perform credential proofs.

1 Introduction

Our modern, fully-interconnected, always-online society relies on a huge number of daily electronic transactions in order to operate, develop and thrive. Internet-based transactions enable e-banking, online commerce and digital communications, while embedded systems and smart cards enable and enhance transactions with public transport, mobile services, electronic purses and TV services.

Authentication methods are required in most of these applications. Although digital credentials provide authenticity, they often include a unique identifier which allows electronic transactions performed by the same user to be linked. Notably, the user may be subject to targeted advertisement without established consent. Racial, medical or personal traits can lead to discrimination and the user’s location may be determined.

In order to address and mitigate the aforementioned privacy threats, several public agencies encouraged advances in credential technology. The European Network and Information Security Agency (ENISA) emphasized the need for “privacy-respecting” use of unique identifiers in European identity cards, the Ontario Privacy Commissioner underlined the necessity of a user-centric approach that embeds privacy into the design and architecture of credential systems [9], while National Institute of Standards and Technology (NIST) issued a strategy towards a user-centric “identity ecosystem” [8]. Last, the European Union promoted legislation (Directive 95/46/EC, General Data Protection Regulation) that enforces the user’s control over his personal data and credentials and establishes privacy by design.

Attribute-Based Credentials (ABCs) emerged as a viable privacy-enhancing technique that implements the new requirements. An attribute based credential is issued to the user by a trusted identity provider. The user’s ABC contains his personal attributes, similar to physical identity documents. The structure of the ABC allows the user to prove to a third party that he possesses a specific attribute without actually revealing the credential identifier. In particular, he is capable of using a zero-knowledge protocol to prove that
he is older than 18 years old, without making his name or identity number public. Moreover, by employing the selective disclosure feature, he can always reveal the minimum required subset of personal attributes in each transaction. Several, cryptographic primitives implement ABC functionality: Microsoft’s U-Prove, based on the discrete logarithm representations (DL-REPs) and the Schnorr signature \[3,5,11\], and IBM’s Idemix, based on the Camenisch–Lysyanskaya (CL) signature \[110\]. The focus of this research is the CL scheme’s existing and desirable functionalities.

Specifically, since only recent efforts try to move ABCs from cryptography to real-world solutions, no general consensus has been reached with respect to the desired properties of the underlying zero-knowledge protocols. Both Schnorr-based and the CL schemes are vulnerable to information leakage (e.g. learning if a user possesses a certain attribute or not), since the selective disclosure procedure can be eavesdropped if an insecure channel is used. The Schnorr-based protocols are more susceptible to identifier leakage, due to the static nature of the signature—as we will see later, it does not provide multi-show unlinkability. Moreover, the current cryptographic basis cannot detect or stop a malicious terminal that impersonates a legitimate one.

These issues can be solved, to a certain extent, by employing an additional layer of authentication and encryption below the zero-knowledge proof, using standard public key cryptography. However, this approach is not efficient when it comes to resource-limited devices such as embedded systems and smart cards. Even in a modern desktop computer, where the computational overhead is small, an additional layer of authentication may leak identifiers. To tackle this problem, we added designation \[4\] to the CL scheme. Instead of the additional layer, we integrated public key cryptography within the zero-knowledge proof. This allows a user to share attributes only with a particular party, making sure that these attributes are revealed only to the designated verifier and does not leak elsewhere. Previous research by Alpár et al. \[2\] has extended the cryptographic basis of DL-REP with the designation feature. Still, the internal structure of Schnorr-based schemes does not allow a credential to be shown more than once as it does not offer multi-show unlinkability. On the other hand, the CL signature is randomized in every transaction, which provides inherent multi-show unlinkability. By adding designation, we can achieve the best of both worlds.

In summary, our research extends the cryptographic basis of the CL signature used in Idemix by adding the designation feature. The resulting zero-knowledge proof is secure assuming the CL signature is secure, which in turn relies on the strong RSA assumption.

2 Technical Background

In this section we introduce attribute based credentials in more detail, and explain the basis of Idemix: the Camenisch Lysyanskaya signatures.

2.1 ABCs and their features

An attribute-based credential (ABC) is a cryptographically protected collection of user attributes. First, an authoritative source, the so-called issuer issues it to a user via an
issuance protocol, and later the user can present it to a relying party, or verifier, via a showing protocol.

![Diagram of ABCs: issuer, user, verifier]

**Fig. 1.** The core modules of ABCs: issuer, user, verifier.

ABCs need to implement specific cryptographic features that provide security for the system and privacy for the users. Specifically, **unforgeability** prevents non-authoritative parties to issue valid credentials. **Non-transferability** ensures that users cannot share their credentials. **Issuing unlinkability** makes sure that the issuer cannot recognize credentials when they take part in showing protocols. **Multi-show unlinkability** renders it impossible for verifiers to link activities of the same user, meaning also that a credential cannot be inherently correlated with its uses.

This paper adds an additional feature to ABCs, namely the **designation** feature. In the context of a zero-knowledge proof used in ABCs, designation integrates public key cryptography inside the proof. As a result, the improved showing protocol can ensure that only the **designated** verifier will be able to verify the user’s credential. Adding the designation property has minimal impact on the aforementioned scheme (Figure 1), since it affects only the showing protocol between the user and the verifier, while the issuing protocol remains unaltered.

### 2.2 Camenisch–Lysyanskaya signature

Existing ABC technologies [10] use the signature scheme proposed by Camenisch and Lysyanskaya (CL scheme) [11]. Its security relies on the strong RSA assumption and it provides efficient zero-knowledge proofs. Before presenting our designated variant, we briefly recall the CL scheme.

The scheme requires a system parameter \( n = pq \), where \( p, q \) are safe primes, generated randomly by the issuer and kept secret. All computations are performed in the quadratic residue \( QR_n \) subgroup of \( \mathbb{Z}^*_n \). The issuer also generates the public constants \( R_0, \ldots, R_l, S, Z \in QR_n \). A basic CL signature on a block of \( l + 1 \) messages is constructed as follows.

**Signature generation:**

\[
A := \left( \frac{Z}{R_0^{m_0} \cdots R_l^{m_l} S^v} \right)^{1/e} \mod n
\]

**Signature verification:**

\[
Z \equiv A^e R_0^{m_0} \cdots R_l^{m_l} S^v \pmod{n}
\]

By knowing the prime factors \( p \) and \( q \), the issuer can compute \( 1/e \pmod{\varphi(n)} \) (\( e \) and \( v \) are random numbers) and then \( A \). The resulting signature on a block of messages
Given a signature and the system parameters, a signature can efficiently be confirmed by the verification equation.

When the CL signature is used for creating attribute-based credentials, \( l \) attributes are represented as numbers \( m_1, \ldots, m_l \) and \( m_0 \) is the user’s secret key. To achieve the desired privacy properties, the credential issuing (signing) and its verification happen in a more intricate way than in the basic signature scheme. As the issuing protocol is in fact a blind signature, the issuer does not learn the user’s secret key \( m_0 \) and the resulting signature \( (e, A, v) \). This provides the issuing unlinkability feature. By using randomization and zero-knowledge-proof techniques while showing a credential, a user can achieve multi-show unlinkability.

### 3 Designated Showing Protocol with CL

In this section we present the new credential showing protocol that integrates the designation property (Table 1). The user proves possession of a credential with \( l \) attributes, while keeping the value of all the attributes secret. This prove is designated to the verifier. We followed similar techniques as Bringer [1] and Alpár et al. [2]. Every verifier has a private-public key pair \((k, V)\). During the showing protocol \( V \) is used to derive a designator \( De \). The verifier is the only one that possesses the secret key \( k \), and thus, he is the only one capable of verifying the zero-knowledge proof.

The designated showing protocol’s public input consists of the CL constants \((Z, S, R_0, \ldots, R_{l-1})\), the RSA modulus \( n \) and the public key \( V \), used for designation purposes. The private input to the scheme contains the corresponding designation private key \( k \) and the secret attributes \( m_0, \ldots, m_l \). After the standard randomization process, the ZK commitment phase commits to the secret attributes (using values \( s, t, \{ w_i \} \)) and generates the designator \( De \). The User sends the commitment \( Co \) and designator \( De \) and upon receipt of the Verifier’s challenge \( c \) he generates the response, based on the secret attributes \( \{ m_i \} \), the values \( s, t, \{ w_i \} \) and the designator exponent \( b \). Following that, the Verifier computes the verification equation using the commitment \( Co \), the designator \( De \) and the secret designation key \( k \).

The verification equation needs to be adapted because the designation value \( De \) would require to take \( k \)-th root. This is not a problem in [24] where the order of the group, which is essential in taking roots, is a public system parameter. In the CL setting, where the strong RSA assumption holds, it is impossible to take roots; therefore, we need to restructure the verification equation.

The primary goal of designation is to make it impossible for any external parties to verify the validity of a proof. This can also be achieved by encrypting the whole conversation using a shared key between the user and the verifier. We note, however, that this might not be efficient – due to the large amount of communication overhead and possible identifier leakage during a prior key setup phase. On the user’s side the computational costs of designation is only one modular exponentiation \( V^b \) when compared with the CL scheme, which is feasible even on devices with limited resources, such as smart cards.
### Table 1.

<table>
<thead>
<tr>
<th>User</th>
<th>Public</th>
<th>Verifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>(m_0): secret key</td>
<td>(Z, S, R_0, \ldots, R_{t-1} \in QR_n)</td>
<td>(k): Verifier’s private key</td>
</tr>
<tr>
<td>(m_1, \ldots, m_l): attributes</td>
<td>(n): RSA modulus, size (l_n) bits</td>
<td>(V = (\Pi_{i=0}^{l-1} R_i)^k)</td>
</tr>
<tr>
<td>Signature ((A, e, \nu))</td>
<td>(V): Verifier’s public key</td>
<td></td>
</tr>
<tr>
<td>(v), size (l_v) bits</td>
<td>(l_0): size of security interval</td>
<td></td>
</tr>
<tr>
<td>(e), size (l_e) bits</td>
<td>(l_H): length of hash function</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Randomization</th>
<th>(A')</th>
<th>(\rightarrow)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(r \in R {0, 1}^{l_0 + l_0})</td>
<td>(\nu = v - er (in Z))</td>
<td>(A' = AS^{-r})</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ZK Proof</th>
<th>(\rightarrow)</th>
<th>(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(t \in R {0, 1}^{l_v + l_0 + l_H})</td>
<td>(\leftarrow)</td>
<td>(c \in R {0, 1}^{l_c})</td>
</tr>
<tr>
<td>(s \in R {0, 1}^{l_v + l_0 + l_H})</td>
<td>({\nu, \nu, \ldots, \nu})</td>
<td>(\leftarrow)</td>
</tr>
<tr>
<td>(w_i \in R {0, 1}^{l_v + l_0 + l_H})</td>
<td>({r_t, r_s, r_{m_0}, \ldots, r_{m_l}})</td>
<td>(\rightarrow)</td>
</tr>
<tr>
<td>(C_0 = A^n S^n R_{m_0} \ldots R_{m_l})</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Security Analysis and Discussion

In this section we discuss the security of the proposed scheme and provide a comparison between the designated CL protocol and the existing designated DL-REP scheme, with respect to the offered security features and performance. The comparison is summarized in Table 2.

Security. The security of the designated CL zero-knowledge proof relies on the security of the original CL scheme which is secure under the strong RSA assumption.

We give an informal description of a security reduction to the security of the showing protocol of the original CL scheme. The additional elements in the transcript of the zero-knowledge proof of the designated scheme are \(De\) and the modified responses \(\{r_{m_i}\}_{i=0, \ldots, t}\). \(De\) is a commitment to a random element \(b\) which is added to each original \(cm_i + w_i\). These elements can easily be added to the original CL scheme by an outside party. Therefore, an
adversary that can break the security of the designated scheme can be used to break the security of the original CL scheme.

The properties of completeness, zero-knowledge, and special soundness hold for the designated CL.

**Designation.** While the techniques applied in the Schnorr/DL-REP schemes and in the CL scheme to make designation possible are similar, there is a surprisingly substantial difference. Consider the validation of credentials in the two designated schemes. We recall that in [2,4], after a proof of an identifier, the verifier has to look it up in a database of all valid identifiers. This additional step and the extra database are not required here. While a DL-REP can be produced by anyone, a CL signature can only be computed by the issuer. Therefore, proving a CL signature immediately proves validity of a credential. Furthermore, since Schnorr identifiers and DL-REPs cannot be randomized like a CL signature, a proof of knowledge directly exposes the identifier. (Essentially, they prove that the user knows the exponent(s) of the public value \( I = g^x \) or \( I = g_0^{x_0} \ldots g_l^{x_l} \) in the Schnorr or the DL-REP schemes, respectively.)

**Issuer-Show Unlinkability.** This property guaranteed by the blind signature produced by the issuer during issuance. The issuer does not learn the secret key nor the resulting signature corresponding to the credential. Therefore, the issuer cannot recognize the credential when it is later shown. As blind signature can be realized in (designated) Schnorr-like schemes as well as in the (designated) CL scheme, they both meet this requirement.

**Multi-show Unlinkability.** As mentioned, the user is capable of fully randomizing the signature before sharing it with a verifier. Thus, it fulfills the requirements for multi-show unlinkability and this property is retained under designation. Designated Schnorr does not provide this feature and it recommends the usage of multiple signature tokens in order to ‘randomize’ each transaction [11], which requires more storage.

**Performance.** One of the main drawbacks of CL schemes is its reliance on RSA groups, which require at least 1024-bit keys to achieve an appropriate level of security. On the other hand, Schnorr/DL-REP/U-Prove schemes operate in groups where the DL assumption holds, which means at least a 1024-bit prime-group size or a 160-bit elliptic-curve group. A U-Prove (prime group) implementation on MULTOS smart cards [6] achieved a showing protocol (five attributes proof of knowledge) less than 0.7 second computation time, while the CL scheme for similar parameters [13] resulted in almost 1.3 seconds. Moreover, the U-Prove results can be even further improved with an elliptic group implementation.

### 5 Conclusions

We proposed a new cryptographic scheme, the designated multi-attribute CL, rendering a user of Idemix capable of revealing his attributes only to a designated verifier, without risking identity leakage or identification. The scheme does not modify the issuing and
<table>
<thead>
<tr>
<th>Property</th>
<th>Designated DL-REP</th>
<th>Designated CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Designation</td>
<td>applied on signature</td>
<td>applied on identifier</td>
</tr>
<tr>
<td>Issuer-showing unlinkability</td>
<td>fulfilled</td>
<td>strongly fulfilled</td>
</tr>
<tr>
<td>Multi-show unlinkability</td>
<td>not fulfilled</td>
<td>fulfilled</td>
</tr>
<tr>
<td>Performance</td>
<td>faster</td>
<td>slower</td>
</tr>
</tbody>
</table>

Table 2. Comparison table between designated Camenisch–Lysyanskaya scheme and designated DL-REP/Schnorr schemes [2].

causes just a slight overhead in the showing protocol. We also examined the security and privacy properties of the suggested scheme and compared it with the existing designated DL-REP protocols. In general, we consider the attribute based credentials to be a viable and secure solution for privacy-sensitive applications which are in-line with technical and legal guidelines for privacy protection. On the other hand, there are still several open issues regarding the desired ABC properties and device deployment. For future work, we suggest establishing a solid set of the required cryptographic schemes and make sure that all of them are compliant with the current regulation. Furthermore, we encourage efforts in standardizing ABC protocols, usage, system deployment and interoperability to assist this nascent technology in becoming mainstream.

References

Security Analysis Of An Implantable Cardioverter Defibrillator

E.Marin[1,2], D.Singelee[2], S.Pollin[3]
UPC, Div. EETAC-TSC, Castelldefels, Spain[1]
ESAT-COSIC, KU Leuven, Belgium[2]
ESAT-TELEMIC, KU Leuven, Belgium[3]

Abstract

Recently, wireless communication interfaces have become an intrinsic part of many modern implantable medical devices (IMD). Even though wireless connectivity has enabled the development of new services within these devices, it also incorporates some unique security and privacy issues. Medical data sent by medical devices could be eavesdropped by an adversary. Moreover, unauthorized commands can be transmitted wirelessly to change the settings of the IMD. This could suppose an important security risk since such an attack could potentially be life threatening to the patient.

Thus, the aim of our work is to investigate whether these attacks can be performed on some commonly used, commercial IMDs. By means of a USRP board and tools available in LabVIEW, we were able to reverse-engineer part of the wireless communication protocol between an implantable cardioverter defibrillator (ICD) and a commercial device programmer, and discovered the key aspects of this protocol.

1 Introduction

Implantable Medical Devices (IMDs) are used to monitor and treat physiological conditions within the human body. Most common IMDs include pacemakers, implantable cardiac defibrillators (ICDs), insulin pumps and automated drug delivery systems. They can be used to remotely monitor the vital signs of patients or even change the therapy by reprogramming the device. Thereby, IMDs can help to manage a broad range of ailments such as cardiac arrhythmia, diabetes or Parkinsons disease.

Nowadays, IMDs typically have a wireless interface to report and/or receive commands from an external device operated by medical staff (e.g., a commercial device programmer to communicate wirelessly to an ICD). These improve many aspects of daily medical control, such as remote therapy management without surgical intervention.

Beside the clear benefits, recent studies have shown that the wireless interface of these devices can be exploited to compromise the confidentiality of the transmitted data, and even worse, to send unauthorized commands to the IMD. This entails an
important security risk, since these unauthorized commands could, for instance, be used to deliver electric shocks to the patient or change the insulin dosage.

In this paper, we studied the communication protocol between a commercial device programmer and an ICD using a USRP board and LabVIEW. This knowledge would allow an attacker to send unauthorized commands to the ICD. This paper consists of five sections. We started with a very short introduction to IMDs and explained the problem statement. Section 2 gives an overview of the setting in which our experiments have been carried out. Section 3 explains our approach to carry out software defined radio attacks. Our most important findings are discussed in Section 4. Finally, Section 5 concludes the paper.

2 General system overview

We will now describe how the device programmer is used to communicate wirelessly to the ICD. Firstly, there is a device in the hospital known as device programmer used by doctors to access the ICD. When the patient arrives to the hospital for the first time, the doctor accesses the ICD through the programming head (also known as magnet), which is part of the device programmer. Next, there is a kind of pairing process via short range communication between both devices. Then, they can communicate via long range wireless communication and do not require using the programming head anymore (unless the pairing key expires). In our proposal we focused on this long range wireless communication.

The communication between devices is bidirectional (from ICD to the device programmer and vice-versa). The communication from the ICD to the device programmer is used mainly to send measurements, while the communication from the device programmer to the ICD is used to send commands. Modern implantable medical devices transmit their messages over a specific band of frequencies, known as MICS (Medical Implant Communication Service). The 402-405MHz frequencies have propagation characteristics conducive to the transmission of radio signals within the human body. In addition, equipment designed to operate in this band can fully satisfy the requirements of the MICS frequency band with respect to size, power, receiver design and antenna performance.

3 Experimental results

Before we could start reverse-engineering the wireless communication protocol and the frame structure of the packets exchanged between the programmer and the ICD, we first had to discover some unknown parameters of the wireless communication. For security reasons, Medtronic, one of the major vendors of IMDs, does not disclose some parameters related with the wireless communication. Hence, the main goal of this work is to find all these unknown parameters such as the transmission frequency, symbol rate, frequency deviation and learn the packet structure. It is important that we find the correct values of these unknown parameters, since even small differences could already translate into large errors when attempting to recover the messages sent by the device programmer.
Firstly, to find the transmission frequency we made a frequency sweep which involves examining the frequency content of a spectral band as a function of time. LabVIEW offers a free tool that lets us scan a given frequency range and which outputs the FFT power spectrum.

After performing several tests, we can observe how a signal appears within the MICS frequency band when the doctor accesses the ICD by means of the device programmer (as is shown in Figure 1). This result corresponds to what we expected to observe, since the website of Medtronic openly publish that all their implantable medical devices use the MICS frequency band to operate.

![Wideband Spectrum](image)

**Figure 1: Transmission frequency used by the device programmer**

The device programmer together with the ICD always choose, before establishing the communication, what they consider the best channel to transmit their packets. This means that the transmission frequency varies. Even more, after they have already established the channel they can, according to some criteria (mainly based on power measurements at both communication sides), change the transmission frequency. Therefore, these devices are continuously monitoring the channel and they can, if necessary, jump to another available channel.

In our experiments, the frequencies that appeared the most were: 403.95MHz, 404.25MHz, 404.55MHz. From this data, we can also extract another important parameter: the channel bandwidth, which corresponds to 300 KHz. According to this, the channels may be allocated in the following frequencies: 402.15 MHz, 402.45MHz, 402.75 MHz, 403.05MHz, 403.35MHz, 403.65MHz, 403.95MHz, 404.25MHz, 404.55MHz, 404.85MHz.

To find the modulation scheme, we first must consider that both devices can be using different modulation schemes. In this paper, we will only focus on the modulation scheme used by the device programmer.

We first looked if the device programmer could be using a frequency modulation scheme. By means of the program used to find the transmission frequency, we can now zoom in on the transmission frequency to see the shape of the signal. Figure 2 shows the extended signal after zooming in on the transmission frequency. We can observe two main peaks in the center of the image (the other peaks correspond to the different signals harmonics).
This can only mean that the modulation scheme used by the device programmer is a BFSK (Binary Frequency Shift Keying), also known as 2-FSK.

When talking about FSK signals, some important parameters must be considered such as the mark frequency , space frequency and frequency deviation. We refer to frequency deviation as the absolute value of the difference between the center frequency and the mark or space frequencies. To find the frequency deviation, we can even zoom in more on the signal (in the frequency domain) to measure the difference between the center frequency and one of the two peaks of the signal (mark or space frequency).

To find the symbol rate, we observe the digital waveform graph after demodulating the signal. This displays the waveform of the signal transmitted by the device programmer as a function of time. Using this graph, we want to discover where a symbol exactly starts and ends.

We did this by creating a known signal and transmitting it using a 2-FSK transmitter we have built in LabVIEW. We then received this signal in our 2-FSK receiver and looked how these bits were represented in the digital waveform graph. This allowed us to find the duration of the symbol. Since this method is not very accurate (because we need to be very precise while placing the markers), we had to repeat the same experiment several times. Figure 3 shows the waveform of the signal transmitted by the device programmer. This figure also displays where we placed the markers to determine the duration of a single symbol.

1By convention, the mark frequency is defined as the frequency to represent 1s, while the space frequency is defined as the frequency to represent 0s.

Figure 2: Modulation scheme used by the device programmer

Figure 3: Waveform of the signal transmitted by the device programmer
After we have found all these parameters, we started analysing the main characteristics of the packet and studied the content of the exchanged messages as well as the transaction time-line. The device programmer itself allows us to perform several types of operations such as changing the patient’s parameters (e.g., name, ID, date of birth, etc.), changing medical therapies (e.g., type of detection, frequency, etc.) or interrogating the ICD. Each of these operations follows a different communications protocol (i.e., number of messages being exchanged, etc.). This approach allowed us to perform several operations (e.g., repeatedly changing the patient’s name) and observe the communication flow.

Figure 4 and Figure 5 respectively show the messages exchanged between both devices in two common situations: When the device programmer does not send any command to reprogram the ICD and also when the device programmer is reprogramming the ICD by changing the name.

At first instance, we noticed that the length of the packets transmitted by the device programmer is always 205 bit, while the length of the acknowledgement sent by the ICD often varies. The length of the latter depends on the packet sent by the device programmer.

Every packet transmitted by the device programmer has a start-of-frame that makes it easier for the ICD to detect the beginning of a packet. Although we expected to find only one start-of-frame, after examining many packets, we could conclude that the device programmer can use four different start-of-frames. However, we also noticed that two of these sequences are the bit complement of the other two. This makes us believe that a differential encoding is being used by the device programmer to encode the data before transmitting.

Even when we used different types of ICDs (Medtronic Virtuoso and Medtronic Secura), we could see that these sequences remained the same. We then checked that
the device programmer does not consider the transmission frequency, serial number or active session to choose the start-of-frame. Lastly, we also attempted to find out whether a certain start-of-frame of a particular packet within the message exchange flow (e.g., the first packet being sent by the device programmer) remained the same while repeating the same test several times. This turned out not be the case.

Regarding to the end-of-frame, it was not as obvious as the start-of-frame case. In this case it was not very clear whether the device programmer used, or not, a sequence to report when a packet ends. Since the ICD knows the length of the packets transmitted by the device programmer, it can use this information to know where the packet ends. Based on our observations, we think that no end-of-frame is being used. Nonetheless, we cannot fully discard the (rather unlikely) possibility that an end-of-frame is being used, but that we could not observe this bit string because of some kind of security mechanism being applied to make it more difficult for an adversary to find this sequence.

Using the start-of-frame and the length of the packet we could filter the noise and save the packets in a file. By repeating the same operation a few times and comparing patterns, we looked for similarities between packets. Mainly, we followed two strategies: Firstly, we compared all the packets after making an operation within the same test (e.g., compare the seven last packets in the protocol executed before the device programmer sends the packet that changes the patient name). Secondly, we captured several times the same packet after making the same operation (e.g., compare the last packet in the protocol executed before the device programmer sends the packet that changes the patient name).

Based in the obtained results in our experiments, we believe that these devices could use some kind of security mechanism to make it more difficult for an attacker to eavesdrop the wireless channel. We have partially discarded the possibility that these devices could be using an encryption technique since we can see some similarities between packets. Hence, we suspect that a kind of data scrambling is being used by the device programmer to send the commands to the ICD. This method does not involve an important computational cost since its aim is to send the same bits but disordered (e.g., the bit number 2 as the bit number 20 and vice-versa).
4 Conclusions

Significant advances in the field of biomedical engineering that have occurred in recent years have enabled implantable medical devices to communicate wirelessly with an external device, known as device programmer. Currently, thousands of patients experience the benefits from IMDs in regulating heart rhythm, controlling blood pressure or administering drug dosages. However, recent studies have revealed security vulnerabilities on existing commercial devices which may cause life threatening consequences.

In our paper, we have attempted to mount software defined radio attacks against an implantable cardioverter defibrillator (ICD) by means of a USRP board and tools available in LabVIEW. We discovered important communication parameters such as transmission frequency, modulation scheme used by both devices, frequency deviation, symbol rate. Moreover, we were able to find important frame parameters, such as the packet length, synchronization sequences, etc.

We have found that a differential encoding is being used by the device programmer. Moreover, based on our experiments, we think that these devices have some kind of security mechanisms to avoid that an adversary could easily eavesdrop the wireless channel between the devices, as this would allow mounting an attack against the ICD. We have partially discarded that these devices are using encryption, since the received data is not really random. Hence, we suspect that a kind of data scrambling technique could be used. Further research is needed to prove this hypothesis.

References

Abstract

Fall incidents and the sustained injuries represent the main causes of accidents for elderly people, and also the third cause of chronic disability. The rapid detection of a fall event can reduce the mortality risk, avoiding also the aggravation of injuries. In this paper an automatic fall detector based on microwave radar measurements is presented. A Continuous Wave (CW) Doppler radar is used to detect the changes in speed of different persons experienced during four activities, namely falling, walking, sitting, and no movement. The measurements, performed with the radar fixed on the wall, are introduced in a machine learning process to estimate an activity classification model. A sliding window principle is then used to detect fall incidents in signals consisting of multiple activities. Experimental results, conducted on real human volunteers in a real room setting, have shown a success rate of 95% in detecting fall events. Moreover, no false positives have been reported.

Keywords: classification, fall detection, signal detection, LS-SVM, signal estimation, sliding window, radar remote sensing.

1 Introduction

The elderly population has been steadily increasing worldwide. This situation, together with the shortage of nursing homes and the natural desire to stay at home, has resulted in a growing need for healthcare approaches that emphasize routine long-term monitoring in the home environment. Elderly people who live alone are usually exposed to health risks which in some cases may cause fatality. In fact, fall incident among the elderly is considered one of the major problems worldwide, and often result in serious physical and psychological consequences [1]. Research pointed out that 30% to 45% of the persons older than 60 years fall at least once a year. People who experience a fall event at home, and remain on the ground for an hour or more, may suffer from many medical complications, such as dehydration, internal bleeding, and cooling, and half of them die within six months [2]. The delay in hospitalization increases mortality risk. Studies have shown that the longer the person lies on the floor, the poorer is the outcome of medical intervention [3-4]. For that reason, it is imperative to detect falls as soon as they occur such that immediate assistance may be provided.

Current health monitoring systems are based on necklace or wristwatch with a button that is activated by the patient in case of an accident. However, in emergency situations, this imposes an important risk factor. In fact, the person may forget to wear the device, or likely may no longer be able to press the button. The ideal solution is therefore a contactless approach that avoids the need for actions by the elderly person. Systems under investigation in the latter category are based on video cameras, floor vibration, and acoustic sensors. In the case of the video camera method, researchers are currently trying to address challenges related to low light, field of view, and image processing, but also privacy is a concern [5]. Floor vibration and acoustic sensors have limited success due to the environmental interference and background noise [6].
Due to the disadvantages of existing fall detection technologies, there is a need for further solutions. An alternative approach based on radar techniques has been demonstrated by the authors [7], [8]. The system uses a machine learning technique to distinguish fall events from normal movements as described in [9].

In this paper an automatic fall detector based on microwave radar measurements is presented. In comparison to [9], where the technique is applied to classify signals consisting of one single activity whose starting and ending points are known, a sliding window is now introduced to estimate and to process signals consisting of multiple activities. The size and the overlapping among sliding windows have been optimized for this application.

In Section 2 the automatic fall detector is presented, and the experimental results are discussed in Section 3.

2 Automatic fall detector

The health monitoring system used to design the automatic fall detector has been described by the authors in [8]. It consists of a sensor, combining both radar and wireless communications features, and a base station for data processing (Fig. 1). The sensor integrates a radar module, a Zigbee module, and a microcontroller, while the base station consists of a Zigbee module, a microcontroller, and a laptop.

A Continuous Wave (CW) waveform at 5.8 GHz is generated and used to detect the speed signals produced by the test persons during four different activities, namely falling, walking, sitting down, and no movement. The resulting baseband signals are digitized and transmitted to a base station to be processed using Matlab.

A movement classification based on a Least Squares Support Vector Machines (LS-SVM) approach combined with Global Alignment (GA) kernel [9] is applied to analyze the digitized baseband speed signals in order to distinguish falls from the other activities. The fall detector aims at assessing the changes in speed experienced during a fall or a normal movement. During a fall, in fact, the speed continuously increases until the sudden moment when the movement stops abruptly. During a normal movement, the Doppler signal experiences a controlled movement. More precisely, while a person is sitting down, the speed first gradually increases, and then decreases to a smooth stop, whereas during a walk, instead, the speed is quite constant over time.

The developed algorithm consists of two stages of data analysis, namely the training phase and the testing phase (Fig. 2). Both phases use the digitized speed signal as input.
A. Training phase

The training phase consists of activity detection and segmentation, feature extraction, feature selection, and model estimation. Two classification models have been estimated for the different types of events. More precisely, the four activities are divided in two main groups, namely fall events and normal movements (i.e., walking, sitting, and no movement). These acquired activities are used to build a data set. However, before learning a model, each activity is grouped in a segment of 2 seconds, considered sufficient to cover the details of the activities and mainly the fall event. This operation of segmentation consists in the detection of the activity energy's peak and in cutting the signal around this peak. Given such segments, the data is preprocessed, namely it is first standardized, such that each dimension has zero mean and unit standard deviation, and then transformed using the Short Time Fast Fourier Transform (STFT) from which only the magnitude spectrum is retained. As opposed to the FFT, the STFT can represent time dependent structures and therefore results in higher performance in case of signals that experience a gradual change in velocity. The STFT is performed by cutting first the segments into 50% overlapping frames which are each multiplied with a Hamming window after which the FFT is computed on each of these frames. Prior to the learning phase, the data is again standardized. Once the learning process is finalized, the model is created and stored in a memory to be used in the validation stage.

B. Testing phase

To validate the classification models, an independent test set, with data not used in the training phase, is needed. For this purpose, the stored test signals consist of multiple activities invoked at unknown instants.

The algorithm performed in this phase presents a structure similar to the data processing of the training phase (Fig. 3). However, the main difference lies in the segmentation stage where the sliding window principle is applied due to the fact that the starting and ending points of the activities are unknown. The size of the sliding window is fixed to 2 sec., to be consistent with the length of the activities’ segments in the training phase, while the overlapping should be optimized in order to be sufficiently adequate to distinguish a fall from normal movements taking into the account the number of required recourses and the computational burden in achieving the algorithm. Experimental tests have demonstrated
that a sliding window of 2 sec. with 75% overlapping is adequate to cover the details of the acquired signals. The segments are then re-preprocessed and subsequently arranged to build the test set that is compared with the classification models.

![Diagram](image)

Fig. 3. Block diagram of the testing phase. In the sliding windows stage, the signal is cut in segments of 2 seconds with an overlapping of 75%.

## 3 Experimental results

A training set containing 50 activities executed by a single test person is used to estimate the activity classification models. The models have consequently been tested on 20 speed signals containing multiple activities acquired on 3 different test persons, whose total durations varied from 10 to 30 seconds. Each of these signals was acquired with a single volunteer in the room at a time, and who had not contributed to the training model. Moreover, each signal contains only one fall event invoked in a random instant. The success rate of the algorithm was calculated as the percentage of detected falls.

The results have indicated that the fall detector was able to detect 19 falls, with a success rate of 95%. Moreover, no false positives have been reported.

Fig. 4 shows a test signal of 10 seconds. It consists respectively of a walking movement in the first four seconds, a no movement, a fall event starting at about 6 seconds, and another no movement. The results show that the fall detector detects the fall when the sliding window intercepts the event.

![Graph](image)

Fig. 4. Speed signal including a fall event (F) and normal movements (N). The frequency of the signal is proportional to the radial velocity of the person during the movement. Each detector box corresponds to a window of 2 seconds, dividing the full signal in 17 windows.
4 Conclusions

This paper shows the feasibility of an automatic fall detector based on machine learning techniques and sliding window’s principle. Speed signals, acquired from a human volunteer, have been used to learn an activity model to distinguish automatically fall events from movements (i.e., walking, sitting down, and no movement). The evaluation was performed on data acquired from different persons that have not contributed to the learning of the activity classification model. Experimental results have shown a success rate of 95% in detecting fall events.

Future research will focus on a larger set of activities and on achieving the validation test in real time.

Acknowledgements

This work was supported by FWO-Flanders and KU Leuven GOA project.

References


Model-free 3D Face Shape Reconstruction from Video Sequences

Chris van Dam Raymond Veldhuis Luuk Spreeuwers

University of Twente *
Faculty of EEMCS
P.O. Box 217, 7500 AE Enschede, The Netherlands

{c.vandam, r.n.j.veldhuis, l.j.spreeuwers}@utwente.nl

Abstract
In forensic comparison of facial video data, often only the best quality frontal face frames are selected, and hence much video data is ignored. To improve 2D facial comparison for law enforcement and forensic investigation, we introduce a model-free 3D shape reconstruction algorithm based on 2D landmarks. The algorithm uses around 20 landmarks on the face and combines the structure information of multiple frames. Model based 3D reconstruction methods, such as Morphable Models, reconstruct a 3D face shape model that is strongly biased towards the average face. Therefore, we don’t use statistical face shape models in our model-free approach. The 3D landmark reconstruction algorithm simultaneously estimates the shape, pose and position of the face, based only on the fact that all images in the sequence are recorded using a single calibrated camera. The algorithm iteratively updates the reconstruction by including new frames, while maintaining the consistency of the reconstruction. We demonstrate the convergence properties of the method reflected in the 2D reprojection error and the 3D error with respect to a ground truth model. We show that the quality of the reconstruction depends on the noise on the landmarks. In a second experiment we show that the method can be used on realistic face shape data with a styrofoam head model.

1 Introduction
One of the unsolved issues in forensic comparison of facial data is the comparison with ‘wild photo’ or video data. Law enforcement services are constrained to work with the case material provided, and unlike researchers, they are not able to use recordings from a controlled environment. Among the most difficult problems of ‘wild photo’ materials are the non-frontal pose of faces and low resolution faces, because often material of overview cameras is used for facial comparison. Automatic face recognition software can only handle 2D facial data under a small pose angle. Faces with larger pose angles result in low recognition scores. At the moment there is no automatic face comparison software available for faces under pose. As a consequence often only the best quality frontal face frames are selected, and hence much video data is ignored.

In the future the pose problem may be solved using 3D video techniques, but the changes in both hardware and procedures will be slow. In the meantime the old 2D equipment will still be used in most cases. So for now law enforcement services are still in need of the ‘tools’ to compare non frontal faces. However, these ‘tools’ should treat

*In cooperation with the Netherlands Forensic Institute.
the video data in such a way that no supplementary information is added to the video data. Reconstruction methods, such as Morphable Models [1], reconstruct a 3D face shape model that is strongly biased towards the average face. Such reconstructions could lead to weird forensic cases, where innocent persons are suspected of a crime because of their average-looking faces. In our method we try to avoid this situation, caused by facial models.

The experiments we perform are comparable to a case where a single ATM-camera has caught the face of a suspect in several frames. Each frame contains a different view on the face of a suspect. We assume that the calibration parameters of the camera, such as focal length, principal point and skew, are available. The recording is assumed to contain different views of a face without variation in facial expression. Our final goal is to reconstruct the face in 3D.

In this paper we introduce a model-free 3D shape reconstruction algorithm based on 2D landmarks, so no additional statistic face models or average face models will be used. We use around 20 landmarks on the face to estimate the shape of the face along with the pose and the position of the face for each view. In our first experiment we use simulated data to demonstrate the convergence properties of the method reflected in the 2D reprojection error and the 3D error with respect to a ground truth model. In the second experiment we show that the method can be used on realistic face shape data with a styrofoam head model. The reconstructed 3D models contain only shape information and no texture information.

2 Background

Our problem, of the face of the suspect moving in front of a static camera, is equivalent to a problem where the camera is moving and the suspect is static. So for each view \( i = 1..N \) we have to find the external camera parameters of that specific view. The static shape of the face can be described by \( j = 1..M \) 3D landmarks. We will use \( M \) 2D landmarks with known correspondences to the 3D landmarks in all \( N \) views to obtain a 3D reconstruction of the landmarks on the face. Our camera is described by the pinhole camera model [2], see Figure 1:

![Figure 1: Pinhole camera model.](image)

where a 3D point \( X \) is projected on the image plane in 2D point \( x \).

Sturm and Triggs provide a method to obtain a projective structure \( \hat{X} \) and projective camera \( \hat{P} \) by factorization of the projections \( x \) of all views [3]:

\[
\lambda_{ij} x_{ij} = \hat{P}_i \cdot \hat{X}_j = P_i \mathcal{H} \cdot \mathcal{H}^{-1} X_j
\]

where \( P_i \) is a \( 3 \times 4 \) camera matrix, containing the calibration, rotation and translation parameters of view \( i \), \( X_j \) is a \( 4 \times 1 \) homogeneous 3D vector of point \( j \), \( x_{ij} \) is a known homogeneous 2D vector of the projection \( i \) of landmark \( j \) and \( \lambda_{ij} \) is a scale factor.
representing the projective depth of \( x_{ij} \). The projective transformation \( \mathcal{H} \) is a \( 4 \times 4 \) transformation matrix, which defines an unknown projective ambiguity between the points \( X_j \) and camera matrices \( P_i \). If the projective depths \( \lambda_{ij} \) are known, the system of equations is of rank 4. The projective depths can be estimated using epipolar geometry on pairs of views, see [3] [4] for details. Singular Value Decomposition (SVD) can factorize the projections, using a rank 4 approximation of the projections, in \( \hat{P}_i \) and \( \hat{X}_j \). For details and properties of SVD see, for example, [5]. The main problem with this factorization method is that all landmarks should be visible in all views, but in realistic cases around 30% of the landmarks on the face are hidden due to the pose of the face. Another disadvantage of this method, for us, is the ambiguity \( \mathcal{H} \), which makes it difficult to merge, for example, multiple partial solutions.

With the calibration parameters available, we can use a method described in [6] to estimate the rotation and translation parameters for one pair of views. This method provides four solutions for the rotation and translation parameters, see 2.2, but only one of these solutions consists of points in front of our camera:

\[
\begin{align*}
\hat{P}_1 &= \begin{bmatrix} UWV^T & +u_3 \end{bmatrix} \\
\hat{P}_2 &= \begin{bmatrix} UW^TV^T & +u_3 \end{bmatrix} \\
\hat{P}_3 &= \begin{bmatrix} UWV^T & -u_3 \end{bmatrix} \\
\hat{P}_4 &= \begin{bmatrix} UW^TV^T & -u_3 \end{bmatrix}
\end{align*}
\] (2.2)

where the rotation matrix defined by \( U \) and \( V \) is the result of a singular value decomposition of the essential matrix, the translation \( u_3 \) is the last column of \( U \) and \( W \) is a matrix that mirrors one of the axis, see [2] [6]. This solution has 5 degrees of freedom, 3 for the rotation and only 2 for the translation, because the equation is determined up to an unknown scale. The rotation and translation parameters are extracted directly from the essential matrix of one pair of views. After estimating the rotation and translation, we can estimate the structure by linear triangulation of the pair of views [2]. This method reconstructs only all visible points in 2 views, but can be extended to more than 2 views. In our case we have a low number of landmarks, so the reconstruction based on two views gives a poor estimation of the shape. Therefore, we extend the algorithm using multiple views to overcome the problems of noise and the low number of landmarks. We introduce an algorithm that iteratively updates the reconstruction by including new views, while maintaining the consistency of the reconstruction for a low number of landmarks.

### 3 Reconstruction Algorithm Overview

**Step 1: Estimate Translation and Rotation Parameters**

Determine the relation between all pairs of views by calculating the essential matrices for each pair. This leads to \( N \times N \) essential matrices, where \( N \) is the number of views. For each pair of views the relative rotation and translation (except for the scale) can be calculated from the essential matrices, see 2.2. The essential matrices have 5 degrees of freedom and therefore require at least an overlap of 5 landmarks for each pair of views in a noise free case. We will use 8 landmarks as a minimum to be able to handle some noise. The essential matrices are estimated with a robust MSAC method (M-estimator SAmple Consensus) [7]. If less than 8 common landmarks are visible between a pair of views, the rotation and translation of this pair are marked as undefined.
Step 2: Estimate Missing Rotation Parameters

Find a solution for missing rotations between a pair of views. Some of the missing rotations can be estimated based on views defined via other views. A new $N \times N \times N$ array is created with rotations defined via intermediate views. So for each reference view, all views via other views are calculated by multiplying the rotations of these views. If one of the three views is not defined, the rotation is undefined. If, for example, the essential matrix between view 1 and 4 was not available, but the essential matrices between view $1 \rightarrow 2$ and $2 \rightarrow 4$ were available, a new rotation can be calculated.

Step 3: Selecting Robust Rotation Parameters

The estimations of the rotations for each pair of views are made more robust, based on the largest cluster of points in the 3D angle space of all possible estimated rotations. The average of the largest cluster is used as a new estimation of the rotation between the pair of views. The cluster should at least contain 5 points to be considered as a robust estimation. Otherwise the original estimation will be used, if available. A new $N \times N$ array of rotations can now be calculated, which contains at least as many estimated rotations as before. Finally the translations of the original views are added to the corresponding rotations. For newly estimated rotations, the translations are set to zero.

Step 4: Selection of Reference View

We select view $k$ with the lowest number of undefined points as reference view for our starting pair of views. We calculate the relative rotations and relative translations to view $k$ for the total $N \times N$ set of rotation and translation estimations.

Step 5: Optimization of Translation Parameters

The translation estimations are optimized by minimizing the reprojection error of all camera estimations. The shape for this minimization is calculated based on view $k$ and the current view. All estimations are relative to view $k$ and can therefore be exchanged to obtain a total set of views with minimum reprojection errors.

Step 6: Selection of Initial Views and Initial Structure

For each available pair of views with view $k$ as reference, an estimation of the structure is obtained. The pair with the most defined views and the lowest total reprojection error is chosen as the start for our iterative optimization. All views with the lowest reprojection errors based on the estimated structure are selected to be the initial estimations of the other views. We have now only $N$ estimations of the rotation and the translation left. Our iterative algorithm starts with the two views used for estimation the structure.

Step 7: Iterative Optimization of All Parameters and Structure

In each round of this step, one new view is added to the selection set. A new estimation of the structure is made based on the current selection set. The reprojection error of all views in the selection is minimized by optimizing the rotation, the translation and the structure. The next candidate view for the selection set is chosen by checking the convergence behavior of each view. The one with the quickest convergence in 10 iterations will be chosen as the next candidate. This process continues until all
views are added and optimized. To prevent overfitting, only 30 iterations are used for each round. Before adding a next view we do a quick optimization of the rotation and translation parameters of the views, which are not in the selection set, using the current structure estimation.

4 Experiments

4.1 Experiment on Simulated Data: Setup

The goal of the first experiment is to determine the influence of the number of views on the reconstruction, and to investigate the convergence properties of our algorithm. We create a random point cloud of 25 3D points and obtain a set of 100 projections of this point cloud with variation in rotation and translation. Later on, the point cloud is used as ground truth for the 3D error. The calibration information and a random selection of the projections are used in the reconstruction algorithm. We performed two experiments in which we add a different level of Gaussian zero-mean noise, with a standard deviation of 1.0 and 2.0 pixels respectively, to the projections. The noise is added independently to the x- and y-coordinates of the projections. Finally we use a random mask to hide 30% of the data to imitate the hidden landmarks on a face. We use our reconstruction algorithm to estimate the 3D point cloud and the external camera parameters of the views. The quality of the reconstruction will be determined by the 2D RMS reprojection error $E_{2D}$:

$$E_{2D} = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \| x_{ij} - \hat{P}_i \cdot \hat{X}_j \|^2}$$ (4.1)

where index $i$ represents a view $1..N$ and $j$ represents a point $1..M$, $\hat{P}$ contains the external camera parameters of each view and $\hat{X}$ contains a collection of homogeneous 3D points. All projections that were not visible, were left out of the equation, so $MN$ is defined as the total number of visible landmarks summed over all views. The parameter $x_{ij}$ describes the known projections with Gaussian zero-mean noise.

After reconstruction the 3D RMS error $E_{3D}$ between the reconstruction and the ground truth point cloud can be calculated with:

$$E_{3D} = \arg\min_{\mathcal{H}} \sqrt{\frac{1}{M} \sum_{j=1}^{M} \| X_j - \mathcal{H} \hat{X}_j \|^2}$$ (4.2)

where $\mathcal{H}$ is a rigid 3D transformation which aligns the ground truth point cloud $X$ with the reconstruction $\hat{X}$ and $j$ is the index of a point. Note that the point cloud is not aligned to a frontal face position, because of the rigid transformation. The 3D RMS error is only calculated if all points were reconstructed by our algorithm, which might not be the case if one of the landmarks is hidden in all views. The experiment is repeated a 100 times with different instances of noise to see whether our algorithm is robust.

4.2 Experiment on Simulated Data: Results

For Gaussian noise on both the x- and y-coordinates with a standard deviation of 1.0 pixels the graphs in Figure 2 show the expected behavior. The more views are added, the more robust the reconstruction is. If the shape is estimated perfectly, then we
would expect the 2D reprojection error to converge to the level of noise added. The 2D reprojection error converges to an asymptote of $\sqrt{2} \approx 1.41$, which is the expected level of noise, see the left graph of Figure 2. Another observation we make is that a number of views above 30 has little influence on both the 2D and the 3D average error. The robustness of the algorithm is only slowly increasing for more than 30 views, see the right graph of Figure 2. So adding more than 30 views seems to have a small impact on both the quality and robustness of the algorithm. Therefore, also computation time and computation cost become interesting at that point.

If the level of Gaussian noise is doubled to a standard deviation of 2.0 pixels, the behavior is similar to the experiment with a standard deviation of 1.0 pixels. The asymptote here is $\sqrt{8} \approx 2.83$, see the left graph in Figure 3. Adding more views has less effect on the robustness of the reconstruction algorithm, but it still has a decreasing effect on the average reprojection error. When more views are added, the average 3D error also decreases slowly, though the robustness of the algorithm seems not to increase. For more than 35 views, the system shows even more variation in the 3D errors than for 35 views, see Figure 3. This could be explained by the fact that the more views are added, the higher the change for heavy outliers in the projections. From previous experiments we learned that the algorithm is known to be sensitive to heavy outliers. Since none of the selected views are skipped, outliers might severely decrease the result of the reconstruction.

These last graphs of this first experiment, see Figure 4, show the number of times the reconstruction has failed. The reconstruction is assumed to be failed, if the reprojection error is above 5.0 pixels. For the experiment with Gaussian noise with a standard deviation of 1.0 pixels and more than 30 views, the algorithm converges to a solution.
in about 99% of the cases. In the case of a standard deviation of 2.0 pixels the algorithm only converges in 75% of the cases. So the algorithm seems stable for Gaussian noise with a standard deviation of 1.0 pixels, but becomes less stable for Gaussian noise with a standard deviation of 2.0 pixels or above.

4.3 Experiment on Styrofoam Head: Setup

The goal of the second experiment is to determine if the algorithm is capable of working with manually labeled face data. We acquired a 3D model of a styrofoam head with 22 markers located on the face. We created 50 renderings of the model with different rotation and translation parameters. The calibration of the camera is known. The landmarks of the projections were labeled manually. In contrast to the previous experiment, no noise was added to the projections, leaving us with only the noise of the manual landmarking. The reconstruction is based on the calibration data and all of the renderings. Similar to the previous experiment, the quality of the reconstruction will be expressed by the 2D RMS error and the 3D RMS error. The ground truth 3D points are manually labeled in the 3D data of the styrofoam head.

4.4 Experiment on Styrofoam Head: Results

The 2D reprojection error of the reconstructed styrofoam head, see the right picture in Figure 5, is 1.96 pixels, which is probably the accuracy of the manual landmarking of the 2D dataset. This noise level is similar to a 1.4 pixels error in both x- and y-coordinates. A rough estimation gives us a head size of 300 mm and the size of the head in the frames is around 500 pixels. So each pixel represents 0.6 mm. Our method is able to estimate the landmarks with $(1.96 \times 0.6 =) 1.2$ mm precision on average. The 3D error is 1.10, which is around 0.6% of the size of the head. The results are in line with the results of the first experiment on simulated data. Our algorithm has
similar convergence properties and errors, and can therefore be applied on manually labeled realistic face data.

5 Conclusion

The experiment on the simulated point cloud shows that the quality of our reconstruction depends on the level of noise in the projections. For small level of noise, around 1.0 pixels, the convergence and robustness of the algorithm seem sufficient. For a larger level of noise the system might become instable, and even not converge to a useful solution. For Gaussian noise with a standard deviation of 2.0 pixels the algorithm only converges in 75% of the cases. The minimum number of views needed to get sufficient quality for the reconstruction is around 30 views. More views can improve the reconstruction, but this will only give a small improvement. In the second experiment, we showed that manual landmarking leads to an error comparable to a Gaussian noise with a standard deviation of 1.4 pixels. This value converted to an accuracy of 1.2 mm for an average face, indicates that the algorithm is capable of working with manually labeled realistic face data.

5.1 Future Work

Currently, we are working on more extensive experiments on the styrofoam data. The next step will contain experiments on real video data. In future work we will include the texture information in the reconstruction to get a full 3D model of the face. We need to find a method to blend the textures of all the visible views. Another improvement would be to get a more dense reconstruction of the face, based on our estimated reconstruction. This can be done by searching for image similarities in local regions around the known landmarks. A more dense reconstruction provides a more accurate description of the facial structure.

References


Secrecy rate versus verification performance in biometric authentication schemes

Joep de Groot\(^1\) and Jean-Paul Linnartz\(^1\)

\(^1\)Signal Processing Systems, Eindhoven University of Technology, Den Dolech 2, 5612 AZ Eindhoven, The Netherlands, tel: +31402473953, e-mail: j.a.d.groot@tue.nl

Abstract

Helper data schemes (HDS) are used in biometric authentication to extract a noise-free authentication string. One objective is to avoid that the stored template leaks little or no information about the users biometrics. A zero leakage (ZL) fuzzy extractor (FE) is often regarded as the one of the most secure solutions. A concern is the limited verification performance, e.g. a large false rejection rate (FRR). This paper studies whether systems with a non-uniform secret or with non-zero can be more attractive. We argue that the secrecy rate, i.e., secret entropy minus leakage, versus FRR is an appropriate performance trade off. By modeling the biometric verification system as a communication channel, we show that a fuzzy extractor (FE) does not achieve capacity.

1 Introduction

Biometrics are a becoming a popular method to verify people’s identity and, more recently, to extract a secret key to perform cryptographic functions. There appears to be a trade-off between extracting the full entropy of the biometric in order to derive the strongest possible key, and on the other hand reliable verification with a low false rejection rate (FRR).

Helper data schemes (HDS) \([1, 2, 3]\) are used to extract noise-free binary authentication strings from these biometric noisy data. These schemes transfer some auxiliary data, commonly referred to as “helper data” \(w\), between enrollment phase and verification phase. This ensures reliable reconstruction \(\hat{s}\) of secret key \(s\), but \(w\) should not reveal excessive information about the secret key.

Early studies expressed the entropy and capacity of such systems and showed that \(\text{FRR} = 0\) can be achieved without loss of extracted key entropy \([4]\). Nonetheless, in practical systems, the introduction of template protection reduces the verification performance \([5]\). Especially for “low quality” biometrics, i.e. with poor signal-to-noise ratio and/or a limited number of independent features, one wishes to minimize the verification error rate. Applying an error correcting code that is able to handle many errors is usually not an option, since this will strongly reduce the number of error-free bits that effectively contribute to the secret.
We focus on two properties of HDSs, namely (i) uniformity of the extracted secret and (ii) zero leakage (ZL) between helper data and the secret. HDSs with a (nearly) equal probability of $1/N$ for each of the $N$ values of the secret are commonly known as a Fuzzy Extractor (FE) [6]. The second property ensures that no information about the secret is contained in the helper data $w$, thus the mutual information $I(S; W) = 0$. An important contribution of this paper is that we study to what extent these two security-related properties can affect verification performance.

A biometric verification scheme can be modeled as a communication channel that transfers information from the enrollment to the verification event. This information theoretic concept models the achievable rate, which is determined by the secret and error probabilities.

This paper shows that generic HDS schemes, thus not necessarily having the FE or ZL property, are an interesting alternative to minimize verification errors. We give an expression for the resulting verification performance when tweaking the system parameters that sacrifice FE or ZL. Finally we illustrate this for Gaussian distributions. Relieving the FE requirement improves secrecy rate, thus FE do not necessarily achieve channel capacity.

2 Biometric verification scheme

Most commonly accepted verification schemes consist of an enrollment and verification phase. In the enrollment phase the prover provides his biometric data $\mathbf{x} = (x_0, \ldots, x_{M-1})$. From this data, the system extracts a secret $\mathbf{s} = Q(\mathbf{x})$ and stores this safely in the hashed form $(h(\mathbf{s} \parallel z), w)$, where $w$ is the helper data, which is generated as $w = g(\mathbf{x})$ and $z$ is the salt. The salt is a system and/or user specific random string to prevent cross-matching between different databases. In the verification phase, the prover provides his (correlated) biometric data $\mathbf{y} = (y_0, \ldots, y_{M-1})$ to prove his identity. All variables, except for the salt $z$, are length $M$ vectors extracted by some means of preprocessing to ensure that the components are (nearly) independent, but not necessarily identically distributed. Independence can be obtained for example by applying a principle component analysis (PCA) to the raw data. This allows us to analyze each dimension separately. In this case, the total leakage in a verification scheme will be a summation of the leakage per dimension.

Leakage elimination has been studied in [7] for secret values that are equiprobable, thus for Fuzzy Extractors. Here, each interval belonging to a secret has been subdivided in equiprobable intervals to define the helper data. Each helper data intervals is repeated for each interval of the secret. This construction yields helper data whose probability is independent of the enrolled secret.

Meanwhile, it has been argued that verification performance highly depends on effective quantization in the analog (continuous valued) domain for biometrics and continuous-valued helper data within the quantization intervals [1, 2]. Leakage can also occur in this domain, which is a concern [3, 8]. We define helper data $w$ as a continuous variable that indicates the relative position of the enrollment feature $x$ within a quantization interval belonging to a secret $s$. We will focus ourselves on the continuous scheme only, since a discrete scheme can be considered a special case of the continuous version. To achieve ZL the scheme has to take into account the probability
density of the features. ZL is achieved in this case by
\[ s = Q(x) \] (1)
\[ w = g(x) = \frac{F_X(x) - F_X(q_s)}{p_s}, \quad s \in \{0, \ldots, N - 1\}. \] (2)
Quantization function \( Q \) uses quantization boundaries \( q_s \) as lower bounds on contiguous quantization intervals. These boundaries determine the probabilities of the secret, hence
\[ p_s = F_X(q_{s+1}) - F_X(q_s), \] (3)
in which \( F_X \) is the cumulative distribution function (CDF) of feature \( x \). In this construction the total number of quantization intervals \( N \) does not necessarily have to be a power of 2.

In the special case of a FE \( p_s = 1/N \), which simplifies the enrollment functions to
\[ s = Q(x) = \lfloor N \cdot F_X(x) \rfloor, \] (4)
\[ w = g(x) = N \cdot F_X(x) - s. \] (5)
For HDS’s with non-uniform secrets as well as for the FE, this yields a continuous helper data \( w \) that reveals no information about the enrolled secret \( s \). In fact, one can exactly reconstruct \( N \) possible \( x \) values, each in a different quantization interval. This reconstruction is given by
\[ x_s(w) = g_s^{-1}(w) = F_X^{-1}(p_s w + F_X(q_s)) \] (6)
and similarly for a FE by
\[ x_s(w) = g_s^{-1}(w) = F_X^{-1}\left(\frac{s + w}{N}\right). \] (7)
For ease of expression we will refer to these points as ‘sibling points’.

Based on the \( N \) sibling points, out of which (only) one is the enrolled secret value, and the verification sample \( y \), the verifier can decide which is the most likely value of the secret \( s \), so
\[ \hat{s} = \arg\max_s p_{s|y,w} = \arg\max_s f_{Y|X}(y|g_s^{-1}(w)). \] (8)
As shown in [8] for symmetric and fading noise, this even leads to a simple comparison with by helper-data based thresholds \( \tau_s \), with,
\[ \tau_{\hat{s}} = \rho g_{\hat{s}}^{-1}(w) + g_{\hat{s}-1}^{-1}(w). \] (9)
Due to the likelihood-based reproduction, this is infeasible to express algebraically. However, it can be evaluated numerically by using the enrollment quantization bounds \( q_s \) and verification thresholds found in Eq. (9) as follows
\[ p_{s,\hat{s}} = \int_{q_s}^{q_{s+1}} \int_{\tau_s}^{\tau_{s+1}} f_{Y,X}(y,x) \, dy \, dx. \] (10)
From this joint probability we can derive the marginals
\[ p_s = \sum_{\hat{s}=0}^{N-1} p_{s,\hat{s}} \quad \text{and} \quad p_{\hat{s}} = \sum_{s=0}^{N-1} p_{s,\hat{s}} \] (11)
that are used to study the performance of our scheme.
3 Optimizing verification rate

3.1 Sacrificing secret-key rate

Some security applications prefer a uniformly distributed key. However, if the application can handle a non-uniformly distributed key this can be an interesting alternative. Most biometric verification schemes, including the one briefly introduced in Section 2, are built on a FE principle. The choice of the quantization intervals determines the probability of the enrolled secrets. In this section, we will show that it can be rewarding not to consider the secret distribution in relation to the verification performance.

A system is difficult to attack when the secrets are equiprobably distributed. This corresponds to the highest entropy of secret $S$ expressed as $H(S)$, which is commonly known as the secret key-rate. In other words, the secret key-rate should be maximized.

On the other hand, a verifier should be able to reconstruct the enrolled secret with high probability based on verification sample $y$ and helper data $w$. From this pair, a verifier deterministically constructs an optimal secret estimate $\hat{s} = \text{Rep}(y, w)$ and therefore

$$H(S|Y, W) = H(S|\hat{S}),$$

which should be as low as possible.

We aim at maximizing the verification effectiveness, expressed as mutual information $I(S; \hat{S})$, which combines

$$I(S; \hat{S}) = H(S) - H(S|\hat{S}).$$

Maximizing over the a priori distribution of the secrets naturally gives the expression for capacity of our verification channel

$$C_v = \max_{p_s} I(S; \hat{S})$$

This construction searches an optimal distribution for the verification channel. However, such an optimization does not guarantee the maximum secret key-rate. In Section 4 we will show, by means of an example with Gaussian distributions, that verification capacity is not achieved with a FE construction.

3.2 Sacrificing zero leakage

A similar construction can be found when considering leakage. Secrecy leakage does not necessarily have to be zero. Security requirements only demand $H(S|W)$ to be sufficiently large to prevent attacks.

We now address the distribution $p_{S|W}$ of the secret conditioned on the public helper data $w$. We assume that the attacker has a full knowledge of the system and is able to exploit any unbalance in the secret probabilities. This results in a lowered entropy of the secret $S$ and is expressed as $H(S|W)$.

The situation for the verifier does not change. Since $\hat{s}$ is still an optimal estimate fully determined by $y$ and $w$ we can write

$$H(S|Y, W) = H(S|\hat{S}, W).$$

The right-hand-side can be considered somewhat double, since $\hat{s}$ is determined by $w$, however we require this in our next step. By combining these two expressions we get
the conditional mutual information
\[ I(S; \hat{S}|W) = H(S|W) - H(S, \hat{S}, W) \] (16)
To find the capacity in this case, we maximize over the conditional probability density functions of \( w \), hence
\[ C_v = \max_{f_{W|S}} I(S; \hat{S}|W) \] (17)

This construction allows us to exchange leakage for an increased verification rate. However, introducing leakage does not guarantee an improved verification rate. It is even possible to find modifications that cause leakage, but do not improve the verification.

4 Results on Gaussian distributions

We analyze the effectiveness of the two proposed trade-offs by assuming Gaussian distributions for biometrics and for noise. Our enrollment sample \( x \) and verification sample \( y \) have unit variance and are correlated. Moreover, we will only study an average number of quantization intervals, namely \( N = 4 \).

4.1 Case 1: scaling quantization regions

We explore whether secret key-rate can be exchanged for increased verification performance by scaling the quantization intervals. Since the quantization intervals determine the probabilities of the secrets (Eq. (3)) this corresponds to a modification of the probabilities as proposed in Eq. (14).

Initially the quantization intervals are chosen to yield a FE. Secondly, we scale by \( \alpha \geq 1 \) to widen the intervals, hence
\[ q_s = F_X^{-1} \left( \frac{s}{N} \right) \Rightarrow q'_s = \alpha F_X^{-1} \left( \frac{s}{N} \right) \] (18)

This choice for this modification is motivated by the notion that most errors are caused by samples in the innermost intervals when assigning more than 1 bit per dimension. For a FE algorithm on Gaussian biometrics, these intervals are the smallest
to deliver the same secret probability from a higher probability density of the biometric. So noise will mostly affect the quantization of biometrics in the inner intervals. A solution would be to widen the enrollment (and the verification) intervals, yet this influences the secret distribution and the secret entropy, as show in Fig. 1(a). By starting at the FE point, i.e. $\alpha = 1$, the error rate can be reduced by increasing $\alpha$. This error reduction, however, comes at the cost of an reduced entropy in the secret.

The resulting rate of our biometric channel shows that there is an initial increase. Figure 2 depicts an example for SNR = 16 dB. Both $H(S)$ and $H(S|\hat{S})$ decrease when incrementing $\alpha$, but initially $H(S|\hat{S})$ decreases faster. The effect can also be noted for other SNR’s, but it is less prominent as seen in Fig. 1(b).

### 4.2 Case 2: translating helper data PDFs

This second case addresses another type of optimization: relieving the zero leakage demand. This can be acceptable if the effective entropy $H(S|W)$ in the secret remains large enough to prevent attacks. In this case, enrollment quantization intervals are left unmodified and only verification intervals are widened. This still results in a high entropy secret, but reduces the error probability during verification. Unfortunately this can only be achieved by leaky helper data, which reduces the effective entropy in the secret.

When altering our enrollment functions we cannot change $Q(x)$ since that will influence the probabilities of the enrolled secrets. Only $g(x)$ can be altered to improve the verification. Consider the reconstruction function Eq. (7), which gives us $N$ possible sibling points for helper data $w$. If we want to widen the verification regions given by $\tau_\hat{s}$ (Eq. (9)) then we have to scale up the distance between the sibling points defining them. Combining the expression for the sibling points, Eq. (7), with the expressions for enrollment, Eq. (4) and Eq. (5), yields

$$x_\hat{s} = F_X \left( \hat{s} + \frac{w}{N} \right) = F_X^{-1} \left( \hat{s} - \left\lfloor \frac{N \cdot F_X(x)}{N} \right\rfloor + F_X(x) \right).$$

(19)

Simply scaling up the expression, as we did for the enrollment intervals, does not work since that will only scale up the inner intervals and therefore causes an increased error probability in the outer intervals. Ideally, we would only like to scale up the first term.
in the argument of $F_X^{-1}$ on the right hand side of above equation. However, this term is not available to the verifier since it only sees $w$, which is a composite of both the first and second term.

The only solution to achieve this scaling is to apply it before the parts are added and do the same during verification for the estimate. Since we like to scale up the first term and not down we have to add an additional $-\beta s$ to the helper data, hence

$$w = N \cdot F_X(x) - s \quad \Rightarrow \quad w' = N \cdot F_X(x) - (1 + \beta)s, \quad \beta \geq 0. \quad (20)$$

The applied translation $\beta$ is transferred from enrollment to verification along with the helper data $w$ to be applied to the estimate of the secret $\hat{s}$. A graphical explanation of the applied translation is depicted in Fig. 4.

![Figure 4: Translation of the helper data PDFs by $-\beta s$ for $s \in \{0, \ldots, N-1\}$](image)

Similar to the first case we witness a reduction of verification errors when translating the helper data by $-\beta s$. However, this also strongly decreases the effective entropy $H(S|W)$ as can be seen in Fig. 3(a). In this case the reduction in the conditional entropy $H(S|W)$, the first term of Eq. (16), is higher than the reduction of the conditional entropy faced by the verifier $H(S|\hat{S},W)$, which is the second term of Eq. (16). This effect is also confirmed by Fig. 3(b), in which we do not encounter an optimum like we saw in the first case. In other words the leakage does more harm than it helps us when applied in this way. However, this does not exclude that other modifications on $g$ exist, that lead to an improved verification performance.
5 Conclusions

We have proposed two methods that optimize the (effective) secret key-rate versus the (FRR) verification performance. Both methods reduce the entropy in the secret to reduce errors during verification. Yet, it appears that a non-uniform distribution of the secret can lead to a higher verification rate. An example with Gaussian distributions showed that the highest verification rate was not achieved for a fuzzy extractor (FE) but for non-uniformly distributed secrets.

We have not yet been able to show that sacrificing the zero leakage (ZL) demand can be useful to lead to an increased verification rate. The presented example shows that although it reduced the error rate, it did not improve verification rate. However, it might be possible to find a different modification for the helper data generating function $g$ that causes less leakage and at the same time helps the verifier to reconstruct the secret.

References


Comparison of Super-Resolution Benefits for Downsampled Images and Real Low-Resolution Data

Yuxi Peng  Luuk Spreeuwers  Berk Gökberk  Raymond Veldhuis
University of Twente
Faculty of EEMCS
P.O. Box 217, 7500 AE Enschede, The Netherlands
y.peng@utwente.nl

Abstract

Recently, more and more researchers are exploring the benefits of super-resolution methods on low-resolution face recognition. However, often results presented are obtained on downsampled high-resolution face images. Because downsampled images are different from real images taken at low resolution, it is important to include real surveillance data. In this paper, we investigate the difference between downsampled images and real surveillance data in two aspects: (1) the influence of resolution on face recognition accuracy, and (2) the improvement of accuracy that can be achieved by super-resolution on these images. Specifically, we will test the following hypotheses: (1) face recognition performance on real images is much worse than on downsampled images, and (2) super-resolution improves the performance of downsampled images more than real images.

Our experiments are conducted using videos from the HumanID database. In each video, the target person’s face is captured while he is walking towards the surveillance camera. We detect the faces in the video frames using a Viola-Jones face detector. Then we select face images of four different resolutions: two low-resolution and two high-resolution. The high-resolution images are used for gallery and generating downsampled images. We perform two types of face recognition experiments. In the first type of experiments, three face recognition methods are evaluated for images with different resolution. The three methods are (1) Principal Component Analysis, (2) Linear Discriminant Analysis, and (3) Local Binary Patterns. In the second type, we apply two super-resolution methods: (1) a model based method and (2) a feature based method on the low-resolution (both real and downsampled) images and then compute the face recognition accuracy.

1 Introduction

Face recognition at a distance is a challenging topic in the face recognition domain. The face images captured at a distance usually are small and of low quality, thus, they are not suitable for most of the face recognition systems which are developed for high quality images. To improve the face recognition performance of these low quality images, one way is to use super-resolution (SR) to enhance the resolution. Some SR methods have been developed specially for face recognition.

In [5], a model for SR is built based on Tikhonov regularization and a linear feature extraction stage. It includes the face features which would be extracted for face
recognition as a prior information. In [6], canonical correlation analysis is used to project high-resolution (HR) and low-resolution (LR) image pairs to a coherent feature space and then radial basis functions is applied to find the mapping between them. A data constraint is developed in [14] to minimize both the distances between the constructed SR images and the corresponding HR images and the distances between SR images from the same class. In [3], multidimensional scaling is used to transform the HR gallery and LR probe images to a common space so that the distance between them approximate the distance between their corresponding HR images. Another SR method based on morphable model is proposed in [13].

The LR images used in most SR papers are downsampled from HR images. Recently, a few researchers also use data from real surveillance cameras. In [14] and [3], the proposed methods are also evaluated using images from the SCface database [4] which contains surveillance quality facial images captured at three different distances. In [7], videos are captured and enhanced with SR methods for face recognition. A real-world outdoor video dataset captured by a PTZ camera is used in [12].

In this paper, we investigate the difference between downsampled and real LR images for face recognition. The experiments are conducted using images from the HumanID database [8]. This database contains videos in which the target person’s face is captured while he/she is walking towards the camera. We select images with four different resolutions from the videos and we also generate LR images by downsampling. Then we conduct both standard face recognition and super-resolution experiments. Three face recognition algorithms (PCA [10], LDA [2] and LBP [1]) and two SR methods (RL/DSR method [14] and NMCF method [6]) are investigated.

The remainder of this paper is organized as follows. In Section 2, hypothesis are proposed regarding the difference between downsampled and real LR images. Section 3 gives a brief introduction to the two SR methods that are used in our experiments. Experimental setup and verification test results are presented in Section 4. Section 5 concludes the paper.

2 Downsampled vs. real low-resolution image

We expect downsampled images have better performance for face recognition than real LR images intuitively. One of the main reasons is that the registration is much poorer for real LR images. Finding the landmarks on HR faces is relatively easy. Downsampled images can make use of these landmarks for registration. But for LR images, the landmarks usually cannot be located precisely. Thus, we make following hypothesis:

(1) Face recognition performance on real LR images is much worse than on downsampled images;
(2) Super-resolution improves the performance of downsampled images more than real LR images.

An additional hypothesis is made for the training approach. Three different training configurations are possible. In the first configuration, the images from the training set are downsampled to the same resolution as the probe images. The second one upsamples the images of the probe set to the same resolution as the training images. In the third configuration, images from the probe set are also upsamples, but the training images are first downsampled to the same resolution as the probe images and then upsampled to the original resolution. We would expect the first and third approaches to perform better than the second one.
3 Super-resolution methods

3.1 RL and DSR method

In [14], Zou and Yuen proposed a SR method for very low-resolution images which is compatible with various face recognition methods. They introduce a data constraint which clusters the constructed SR images with the corresponding HR images.

Given a set of HR and LR image pairs \( \{ (I_h^i, I_l^i) \}_{i=1}^N \), the relation \( R \) is modeled as

\[
R = \arg \min_{R'} \sum_{i=1}^N \| I_h^i - R'I_l^i \|^2.
\] (1)

This is called relationship learning (RL). It minimizes the distance between HR and the space of LR projected by \( R \).

To acquire better results in face recognition, identity information about the subject is used based on RL. This is called discriminative super-resolution (DSR). A second term is added to (1), see (2).

\[
R = \arg \min_{R'} \frac{1}{N} \sum_{i=1}^N \| I_h^i - R'I_l^i \|^2 + \gamma d(R').
\] (2)

where \( \gamma \) is a constant to balance the two terms. We set \( \gamma \) to 1 in our experiments. \( d(R') \) is represented as

\[
d(R') = \frac{1}{N(\lambda_i = \lambda_j)} \sum_{\lambda_i = \lambda_j} \| I_h^i - R'I_l^i \|^2 - \frac{1}{N(\lambda_i \neq \lambda_j)} \sum_{\lambda_i \neq \lambda_j} \| I_h^i - R'I_l^j \|^2.
\] (3)

where \( \lambda_i \) is the class label of \( I_i \). This makes sure the reconstructed HR images are clustered with the images from the same class and far away from those from other classes.

In the testing phase, for both RL and DSR, we first apply \( I_{SR} = RI_{input} \) to a given LR image \( I_{input} \) to obtain SR image \( I_{SR} \), and then use \( I_{SR} \) for face recognition.

3.2 NMCF method

In [6], Huang and He proposed a SR method where canonical correlation analysis (CCA) is used to project the PCA features of HR and LR image pairs to a coherent feature space. Radial based functions (RBFs) are then applied to find the mapping between the HR/LR pairs. This method finds nonlinear mappings on coherent features. Thus, we will refer to this method as NMCF method in this paper.

In the training stage, firstly PCA features of HR and LR image pairs are extracted to reduce computational costs. Next, these PCA features are projected to a coherent feature space by CCA where the correlation between HR and LR features is maximum. This provides better condition for finding the mappings in the next step. Let \( \hat{X}^H \) and \( \hat{X}^L \) be the PCA features of HR and LR subtracted by the mean, and \( C_{11} \) and \( C_{22} \) be the within-set covariance matrices, and \( C_{12} \) and \( C_{21} \) be the between-set covariance matrices. Compute \( R_1 = C_{11}^{-1}C_{12}C_{21}^{-1} \) and \( R_2 = C_{22}^{-1}C_{21}C_{11}^{-1} \). The coherent features of HR and LR images are

\[
C^H = (V^H)^T \hat{X}^H, \quad C^L = (V^L)^T \hat{X}^L.
\] (4)

\( V^H \) and \( V^L \) comprises eigenvectors of \( R_1 \) and \( R_2 \) when their corresponding eigenvalues are sorted in descending order.
Then RBFs are applied to approximate the mapping between HR and LR coherent features. The function approximation is represented as $C^H = W \Phi$ where $W$ is a weighting coefficient matrix and $\Phi$ is a multiquadratic basis function (see [6] for details). Thus, $W$ can be solved by $W = C^H(\Phi + \tau I_{id})^{-1}$. $\tau$ is a small positive value and $I_{id}$ is the identity matrix.

In the testing stage, for a given LR probe image $I_p$, first compute the coherent features $c^p$ and then apply the learnt mapping to obtain the SR features by

$$c_{SR} = W \cdot [\phi(\|c_1 - c^p\|) \ldots \phi(\|c_N - c^p\|)]^T$$

where $\phi(\|c_i - c_j\|) = \sqrt{\|c_i - c_j\|^2 + 1}$. The SR features are fed to the nearest neighbor classifier together with the coherent features of HR gallery images for recognition.

4 Experimental results

Our experiments are conducted on the HumanID database [8]. We use the videos in which the target person’s face is captured while he is walking towards the surveillance camera, see Figure 1. The faces from the videos are detected by the Viola-Jones face detector [11]. For each video, we select four face images with different resolution: 70×70, 50×50, 30×30 and 25×25. Thus, we have about 400 images for each resolution. The images with resolution 70×70 are used as gallery images. The images with the remaining three resolutions are used as probe sets. We also generate images with different resolutions by downsampling images of resolution 50×50. Images with six different resolutions are generated: 30×30, 25×25, 20×20, 15×15, 10×10 and 5×5. To train the face recognition classifier and the super-resolution system, we use downsampled images from the FRGC database [9]. We use the eye coordinates for registration of the face images. The FRGC database provides eye coordinates of its images while we manually click on the eyes for the images from the HumanID database.

4.1 Face recognition experiments

In this section we provide the face recognition performance of PCA, LDA and LBP on both downsampled and real LR face images. Distance measures employed in 1-nearest neighbor classifier for PCA, LDA and LBP are L1 norm, cosine angle and Chi square, respectively. Three different training configurations are applied for our face recognition experiments:

Config. 1: the images of the training set are downsampled to the same resolution of
the probe images; Config. 2: the training images are with resolution 70×70, and the probe images are upsampled to resolution 70×70 using bicubic interpolation. Config. 3: the images of the training set are first downsampled to the same resolution of the probe images, then upsampled back to the resolution of 70×70 using bicubic interpolation. The probe images are also upsampled to the resolution of 70×70.

Config. 2 and 3 are also the baseline configurations for SR. Because LBP does not require a training set, we only apply Config. 1 and 2 for LBP.

![ROC curves](image1)

**Figure 2**: ROC curves of PCA results, downsampled vs. real: (a) downsampled, Config. 1; (b) downsampled, Config. 2; (c) downsampled, Config. 3; (d) real, Config. 1; (e) real, Config. 2; (f) real, Config. 3.

![Face recognition results](image2)

**Figure 3**: Face recognition results, GAR@FAR=0.1: (a) PCA, (b) LDA, (c) LBP.

We conducted verification experiments. The ROC curves of PCA is shown in Figure 2. We also compute the genuine acceptance rates (GAR) when false acceptance rates (FAR) equal to 0.1 for PCA, LDA and LBP, see Figure 3.

As we can see, resolution changes do not have much influence for downsampled images, but for real LR images, the recognition results decrease significantly as the
resolution becomes lower. At resolution $30 \times 30$ and $25 \times 25$, the GARs of downsampled images are almost the same as for resolution $50 \times 50$, but the GARs of real LR images are much lower. The performance of different face recognition classifiers are of similar trend while LBP is more sensitive to resolution changes than PCA and LDA. Different training configurations influence little when the image resolution is higher than $10 \times 10$.

4.2 Super-resolution experiments

In this section, we apply the SR methods that were introduced in Section 3 to explore the benefits of SR. Images from the FRGC database are used to train the SR system. The training configurations are similar to Config. 2 and 3, but RL and DSR methods are used for upsampling LR images instead of bicubic interpolation. The NMCF method is only designed for Config. 2. The ROC curves of RL and DSR methods with PCA in Config. 2 are shown in Figure 4. The GARs when FARs equal to 0.1 are presented in Figure 5.

![Figure 4: ROC curves of SR results using RL/DSR methods with PCA, Config. 2: (a) dowsampled, RL; (b) dowsampled, DSR; (c) real, RL; (d) real, DSR.](image)

As we can see from these results, all of the SR methods tested in the experiments have no benefits for verification. LBP results drop dramatically after SR. For PCA and LDA, the RL and NMCF methods keep the verification performance at the same level of bicubic interpolation. But the DSR method makes the results worse. The two training configurations also give similar results. To better explain these results, we shown some reconstructed SR images in Figure 6.

Firstly, compare the real LR images with downsampled images at the same resolution, it is easy to find that the downsampled images are clearer and contain more details than the real images. This explains why downsampled images have better performance in LR face recognition. Secondly, the SR images show more details about the faces but also add some artifacts on them. These artifacts are also a problem for face recognition and it may cause the failure of LBP for recognition. Thirdly, as the resolution become lower, the identity information in the images become less, so the SR images become more look like an average face other than the person himself.
Figure 5: Super-resolution RL/DSR/NMCF results, GAR@FAR=0.1: (a) PCA, Config. 2; (b) PCA, Config. 3; (c) LDA, Config. 2; (d) LDA, Config. 3; (e) LBP results; (f) NMCF method. (ds=downsample)

Figure 6: Reconstructed SR images by RL method from (a) real LR images, resolution 50×50, 30×30 and 25×25; (b) downsampled images, resolution 30×30, 25×25 and 20×20; (c) downsampled images, resolution 15×15, 10×10 and 5×5.

5 Conclusion

We evaluated the difference between downsampled and real LR face images for both standard face recognition and super-resolution. Our results show that face recognition on downsampled images performs much better than real LR images. The face
Recognition accuracy hardly decreases when the resolution change is within a certain range. But for real LR images, the verification results drop significantly as the resolution become lower. Moreover, the single image super-resolution methods do not have benefits for face recognition in our experimental configurations. One possible solution is to apply SR method which can make use of the information of multiple images.

References


Distributed Compressed Sampling Architecture for Maximum Likelihood Signal Detection

J. Bodart*, J. Verlant-Chenet*, A. Bourdoux†, J.-M. Dricot*, P. De Doncker*, L. Lampe‡, F. Horlin*
*Université Libre de Bruxelles - OPERA department
jbodart@ulb.ac.be
†Interuniversity Microelectronics Centre, Wireless Research Group
‡University of British Columbia - department of Electrical Engineering

Abstract
Cognitive radios are a new technology introduced to resolve the spectrum scarcity problem by superimposing new services in the already allocated bands under a non-interference constraint. It has been recently demonstrated that the challenging implementation of the signal detectors can be facilitated by using the theory of compressive sampling. In this paper, we consider a distributed network of secondary nodes that cooperate to detect the primary signals. Each secondary node samples the signal periodically at a rate much smaller than the Nyquist rate. The delays inherent to the propagation channel are used to implement a periodic non-uniform sampling detector when the secondary nodes combine their observations. We demonstrate that the proposed detector can efficiently detect the primary user signal, even under fading channels.

I. INTRODUCTION

Many studies have shown that static frequency allocation for wireless communication systems is responsible for the inefficient use of the spectrum. Cognitive radios (CR) resolve the problem by first detecting the still available frequency resources and then using them for their own transmission. Efforts are being made to develop efficient signal detectors capable of scanning a wide bandwidth with a large dynamic range [1]. Recently the theory of compressed sampling (CS) has received an increasing attention as it may help in relaxing the constraints on the design of the CR system [2]. It is based on the fact that a frequency sparse signal may be sampled at a rate significantly lower than the Nyquist rate without losing information. This may potentially facilitate the implementation of the analog-to-digital (A/D) converters and digital processors.

In the CS framework, signal recovery is classically achieved through expensive algorithms. The signal reconstruction may be expressed as a problem of $\ell_1$-norm minimization and typically solved with linear programming algorithms [3]. Other methods such as the orthogonal matching pursuit algorithm have been proposed [4]. However signal reconstruction is not necessary in many signal processing applications, as the designer is only interested in solving an inference problem. Davenport et al. demonstrate that it is possible to tackle the problem of the detection of a known signal in noise or the classification of signals directly in the compressive domain without first resorting to full signal reconstruction [5], [6]. We have recently extended [6] to the optimal maximum-likelihood (ML) detection of linearly modulated signals of unknown parameters in a set of of predetermined subchannels [7].

On the other hand, the use of distributed spectrum sensing algorithms is recommended to cope with the fading phenomenon present in all wireless communications systems. Distributed algorithms rely on the exchange of the observations made at the spatially distributed nodes on a control channel of capacity unfortunately limited in practice. When the observed
signal at each node is sparse, it is advantageous to first compress the information using the CS theory before communicating it to the coordinator node. The existing contributions in the literature aim at reconstructing the wideband signal spectrum at the coordinator node by defining a sparsity model common to all sensing nodes [8], [9].

In this paper, we focus on the detection of primary signals directly in the compressive domain (as in [7]). Therefore the proposed scheme does not necessitate the signal reconstruction as in [8], [9]. We demonstrate that the CS architecture can advantageously be implemented by a distributed network of sensing nodes, each sampling the signal periodically at a low rate. Because the propagation delays to reach the sensing nodes are inherently different, the distributed network can be viewed as a periodic non-uniform sampling architecture when the signals are combined at the coordinator node.

II. SYSTEM MODEL

Fig. 1 describes the baseband system model. It is assumed that the overall bandwidth is divided in a set of $M$ uniformly spaced frequency bands. The primary system is composed of one base station transmitting a signal on one of the subbands. The secondary system is composed of a set of $Q$ cooperative sensing nodes that send their observations to a coordinator node to detect which band is occupied. The transmitter transmits a finite sequence of symbols $I[n]$ of length $N$ ($n = 1 \ldots N$). The complex symbols are supposed independent and identically distributed (i.i.d.) of variance $\sigma^2$. The symbol duration is denoted $T_{symb}$. This sequence of symbols is lowpass filtered by the pulse-shaping filter $g(t)$ (eg. halfroot Nyquist). The signal is then shifted in the frequency domain to the frequency $\Delta f$. The transmitted signal is therefore:

$$x(t) = \sum_{n=1}^{N} I[n] \ g(t - nT_{symb}) \ e^{j2\pi \Delta ft}. \quad (1)$$

Each secondary received signal is affected by a different propagation delay $\tau_i$. The signal at each receiver is multiplied by a coefficient $\alpha_i$ modeling the fading channel. A phase
shift, modeled by the parameter $\varphi_i$, affects the signal at each receiver. Before sampling at the receiver, the signal is lowpass-filtered with an ideal filter $f(t)$ of bandwidth $1/T_s$ where $T_s = T_{symb}/M$ is the sampling period required to satisfy the Nyquist criterion. The signal is corrupted by the addition of an additive white Gaussian noise (AWGN) $w_i(t)$ of variance $\sigma_w^2$. The signal received at node $i$ is:

$$r_i(t) = \alpha_i e^{j\varphi_i} x(t - \tau_i) * f(t) + v_i(t),$$

where $v_i(t) = w_i(t) * f(t)$ is the noise $w_i(t)$ filtered by the ideal lowpass filter. We assume that it is periodically sampled at the symbol rate $1/T_{symb}$ (and not at the Nyquist rate $1/T_s$):

$$r_i[m] = r_i(t = mT_{symb}),$$

where:

$$g_i[n] := g(nT_{symb} - \tau_i).$$

Because a finite number of symbols is transmitted and assuming that the shaping pulse is of finite length $L$, an equivalent matrix model can be built:

$$r_i = H_i I + v_i,$$

where:

- Vector $I$ is composed of the $N$ transmitted symbols.
- Vectors $r_i$ and $v_i$ are composed of the $N + L - 1$ received and noise elements.
- Matrix $H_i$ is defined as the product $H_i = \alpha_i \Phi_i G_i$ where $G_i$ is the matrix (of dimension $N + L - 1 \times N$) representing the convolution with the pulse shaping filter, including the time shift $\tau_i$. The element $(j,k)$ of $G_i$ is given by:

$$G_i(j,k) = g_i[j - k + 1].$$

Matrix $\Phi_i$ performs the shift to frequency $\Delta f$. The element $(j,k)$ is given by:

$$\Phi_i(j,k) = \delta_{jk} e^{j(2\pi \Delta f (kt_{symb} - \tau_i) + \varphi_i)}.$$

The received vectors are communicated to the coordinator node where the combined received vector $r$ is composed:

$$r = [r_1^T \ r_2^T \ \cdots \ r_Q^T]^T.$$

### III. LIKELIHOOD FUNCTION

The occupied frequency band is estimated by using a Maximum Likelihood Frequency Estimator (MLFE):

$$\hat{f} = \arg \max_{\Delta f} f(r | \Delta f).$$

The distributed detection relies on the pre-estimation of the signal delays, the phase shifts and fading coefficients that are therefore assumed to be known in this study. The relation (10) can be rewritten as:

$$\hat{f} = \arg \max_{\Delta f} \mathbb{E}[p(r | \Delta f, I)]$$
where \( E[x] \) stands for the mathematical expectation of \( x \). In the expression (6) of the received signal, the first term is deterministic; only the second term is random and has a Gaussian probability density function (PDF). Thus, \( \bar{r} \) is a Gaussian random variable of mean \( \begin{bmatrix} H_1 I & H_2 I & \cdots & H_Q I \end{bmatrix} \) and variance \( \sigma_w^2 I_P \) where \( I_P \) denotes the size \( P \) identity matrix. We obtain:

\[
p(r|\Delta f, I) = C \exp \left( \frac{-1}{2\sigma_w^2} \sum_{i=1}^{Q} (r_i - H_i I)^H (r_i - H_i I) \right)
\]

\[
= C \exp \left( \frac{-1}{2\sigma_w^2} \sum_{i=1}^{Q} [r_i^H r_i - I^H y_i - y_i^H I + \lambda_i] \right)
\]

where:

- \( C \) is a constant independent of \( \Delta f \) and \( I \);
- \( r_i^H r_i \) is the correlation of the observation independent of \( \Delta f \) and \( I \);
- \( y_i = H_i I \) is the vector that results from the application of a matched filter to \( r_i \);
- \( \lambda_i = I^H H_i^H H_i I \) is approximately equal to the constant \( N \sigma_w^2 |\alpha_i|^2 \) when \( N \) is sufficiently large and if the pulse shaping filters are normalized.

The PDF defined in (12) is thus approximately given by:

\[
p(r|\Delta f, I) = C' \Omega (\Delta f, I)
\]

where \( C' \) is a constant and:

\[
\Omega (\Delta f, I) := \exp \left( \frac{1}{2\sigma_w^2} \sum_{i=1}^{Q} (y_i^H I + I^H y_i) \right).
\]

The exponential in (14) is expanded as a Maclaurin series in order to be able to average the function over the symbols:

\[
\exp (\lambda) = \sum_{\ell=0}^{\infty} \frac{\lambda^\ell}{\ell!}.
\]

**IV. FREQUENCY ESTIMATOR**

The ML criterion reduces to the simplified expression hereafter:

\[
\hat{\Delta} f \approx \arg \max_{\Delta f} E[I] \left[ \Omega (\Delta f, I) \right]
\]

\[
\approx \arg \max_{\Delta f} E[I] \left[ \exp \left( \frac{1}{2\sigma_w^2} \sum_{i=1}^{Q} (y_i^H I + I^H y_i) \right) \right]
\]

\[
\approx \arg \max_{\Delta f} E[I] \left[ \sum_{\ell=0}^{\infty} \frac{1}{\ell!} \left( \frac{1}{2\sigma_w^2} \sum_{i=1}^{Q} (y_i^H I + I^H y_i) \right)^\ell \right]
\]

\[
\approx \arg \max_{\Delta f} \sum_{\ell=0}^{\infty} \frac{1}{\ell!} \left( \frac{1}{2\sigma_w^2} \right)^\ell E[I] \left[ \left( \sum_{i=1}^{Q} y_i^H I + I^H y_i \right)^\ell \right].
\]

To compute the expectation, we use the binomial theorem that states:

\[
(x + y)^n = \sum_{\xi=0}^{n} \binom{n}{\xi} x^{n-\xi} y^\xi.
\]
Thus:
\[
\Delta f \approx \arg \max_{\Delta f} \sum_{\ell=0}^{\infty} \sum_{\xi=0}^{\ell} \frac{1}{\ell!} \left( \frac{1}{2\sigma_w^2} \right)^{\ell} \left( \xi \right)^{\ell-\xi} \cdot \mathbb{E} \left[ \left( \sum_{i=1}^{Q} y_i^H I \right)^{\ell-\xi} \left( \sum_{i=1}^{Q} I^H y_i \right)^{\xi} \right].
\]  (18)

Only the terms where \( \left( \sum_{i=1}^{Q} y_i^H I \right) \) and \( \left( I^H \sum_{i=1}^{Q} y_i \right) \) are elevated to the same power are different from zero because the symbols are i.i.d. by hypothesis and:
\[
\mathbb{E}[I[n]^\gamma] = 0 \quad \forall \gamma \in \mathbb{N}.
\]  (19)

This happens only for even orders \( \ell \) and for \( \xi = \frac{\ell}{2} \). We have:
\[
\Delta f \approx \arg \max_{\Delta f} \sum_{\ell=1}^{\infty} \frac{1}{(\ell!)^2 \sigma_w^2} \left( \sum_{i=1}^{Q} y_i^H I \right)^{\ell} \left( \sum_{i=1}^{Q} I^H y_i \right)^{\ell}.
\]  (20)

(21)

The dominant terms in (21) may be much larger than the second order. Fortunately, the terms in (21) are monotonically increasing with the terms that compose the second order as \( \ell \) grows, since each term is composed of a sum of powers of those terms and only terms that are redundant with others are subtracted from the sum. Thus, it is equivalent to keep only the second-order term in the objective function. Thus, the function to maximize reduces to:
\[
F(\Delta f) := \left| \sum_{i=1}^{Q} y_i \right|^2 = \sum_{n=1}^{N} \left| \sum_{i=1}^{Q} y_i[n] \right|^2.
\]  (22)

The sequence \( y_i[n] \) is obtained by applying a matched filter on each of the candidate bands:
\[
y_i[n] = \alpha_i^* e^{-j(2\pi \Delta f(nT_{\text{symb}} - \tau_i) + \varphi_i)} \sum_m r_i[m] g_i^*[m - n].
\]  (23)

The signal detector reduces to a matched filter applied independently at each sensing node at the symbol rate. The results are transmitted to the coordinator node where the sum is taken to obtain the final metric based on which the occupied band is estimated.

V. PERFORMANCE ANALYSIS

The subchannel estimation error probability is assessed numerically. A 100 MHz bandwidth is uniformly divided in \( M = 20 \) subchannels. A sequence of 50 symbols is transmitted at the rate 5 Msps on each subchannel. The symbols are shaped with a halfroot Nyquist filter of roll-off factor equal to 0.2. In our simulations, 5000 realizations of the Rayleigh fading channels have been generated.

Fig. 2 illustrates the probability of error as a function of the signal-to-noise ratio (SNR) in three scenarios: (i) the ideal channel, (ii) a common Rayleigh channel obtained by assuming the nodes are co-located, (iii) the distributed Rayleigh channel considered in this paper. Rayleigh fading incurs a performance loss compared to the ideal channel. In the case of a distributed sensing network, the system benefits from spatial diversity explaining the performance gain. Fig. 3 illustrates that the performance improves significantly when the number of sensing nodes is increased thanks to the increasing average sampling frequency and the additional source of diversity.
VI. CONCLUSION

This paper investigates the signal detection in a distributed network. The network is composed of a set of sensing nodes that sample the signal at a low rate and send their observations to a coordinator node. The propagation delays inherent to the distributed network of sensing nodes are used to form a global compressive sampling architecture. Contrary to state-of-the-art solutions, the proposed signal detector works directly in the compressive domain and therefore does not require the reconstruction of the received signal. Performance results show that the system benefits from the spatial diversity present in the distributed network. Future work will investigate the implementation of the detector in a real-life testbed (estimation of the Rayleigh fading coefficients and delays...).

REFERENCES

Selecting feature-based models

Amir Jalalirad*, Tjalling Tjalkens*, Jean-Paul Linnartz*

*Eindhoven University of Technology,
Dept. Electrical Engineering, Signal Processing Systems Group
{a.jalalirad, t.j.tjalkens, j.p.linnartz}@tue.nl

Abstract

In a classification problem, we would like to assign a model to the observed data using its features. If the number of these features is large, considering dependencies between them to come up with an appropriate model becomes a challenge. In this paper an efficient method is introduced for selecting a model that fits the observed data sequence with a large number of features. This method is constructed by a modification of an earlier studied algorithm. It is shown that the model selected by this method has the Maximum Likelihood probability and can capture the dependencies between the features of the data sequence. We prove that through this method, the algebraic workload required for calculating the Maximum Likelihood model is reduced.

1 Introduction

There are many real life applications of feature-based classification, including spam email filtering, disease diagnosis, document classification, quality control in factories. A critical intermediate step is to categorize the acquired data into a number of classes based on the fact that they either contain some combination of specific properties or not. This paper proves that the classification success rate can be improved by appropriately exploiting dependencies among these properties. A well-know issue is that the number of properties, which are the features for classification, is often very large, so also is the number of possible models that exploit dependencies. If the number of choices for the models increases, its selection becomes a challenge. Hitherto this is often done in an ad-hoc fashion [1].

In [2] a method is introduced to estimate the probability of a sequence of observed objects. It calculates efficiently the probability of the observed sequence using a mixture over all possible models. This method comes up with an estimate for the probability of the sequence but not an explicit model for the data. Therefore the dependency between the features which can give much insight into the structure of the data cannot be determined. In this paper, we extend that method to an efficient Maximum Likelihood (ML) model selection algorithm to select the features which are more likely to be dependent. This method will produce a model that reveals dependency between some, none or all of the features of the observed objects. Moreover, the method also works for multi-valued features. We also compare this to the Naïve Bayes method [3] which assumes that all the features are conditionally independent. With this assumption, the number of the parameters of the model is reduced to minimum but the resulting simple model may not be able to capture the data complexities. Nonetheless this method is used most frequently because of being simple to implement [4]. The method introduced in this paper can result in the Naïve Bayes model in case of simpler data structures or a model with more parameters in case of more complex data structures. Finally, by using the Bayesian Information Criterion (BIC), we will show that the model selected by our method can produce the observed sequence with a high probability while it is not the most complex model.
2 An example

As an example we consider feature-based classification which is used frequently in optical recognition of handwritten characters [5]. Each handwritten character that is going to be classified into the alphabet letters or digits is considered as an object. The picture in which the character is written is divided into several segments as shown in Figure 1. Then the number of black pixels in each segment is counted. Each segment is a feature of that character object and the number of black pixels in that segment is the value of the feature. According to the values of these features a character is classified as a specific letter/digit.

![Fig. 1. (a) 10 objects with 16 segments (b) probably independent regions](image)

The Naïve Bayes model assumes each of these segments to be independent. But obviously letters in the alphabet are contiguous ink depositions, so the number of black pixels in the neighboring segments are likely to be related. As an example, in Figure 1(b) we can define some regions of dependent segments, like the middle regions that can show some specific curvature in a digit. We can also recognize some other regions on the sides that have independent segments from the middle ones. Consequently, if we consider some of these segments to be dependent, then we have chosen a more complex model that may be able to classify the data better. But we will have a class of models, each of them assuming dependency between some selected segments. Even the Naïve Bayes model is one of the members of that class which simply assumes no dependency between segments. The network method introduced in this paper efficiently selects a model out of this class that finds the right dependencies between the segments.

3 Notations and some definitions

We show sequences or vectors with capital letters in Bold. When necessary, their length is noted by a superscript. Their index in the sequence/vector is shown in braces in the subscript. Throughout this paper we consider finite length of $k$ for feature vectors, $F = (f_1, f_2, \cdots, f_k)$. We show the feature vector index set $\{1, 2, \cdots, k\}$ by $\mathcal{F}$. Following that, we define $S = \{s_1, s_2, \cdots, s_g\}$ as a set of partitioning subgroups of the feature vector indices such that

$$\forall \ 1 \leq i \leq g : \ s_i \subset \mathcal{F}$$

(1)

$$\forall \ 1 \leq i \leq g \ \text{and} \ \forall \ 1 \leq j \leq g : \ i \neq j \iff s_i \cap s_j = \emptyset$$
The features in each subgroup are dependent. Full partitioning of the feature vector will be discussed whereas sub-partitioning will not introduce major changes to our method. See also [2] where sub-partitioning is discussed in detail. Therefore

\[
\bigcup_{i=1}^{g} s_i = \{f_1, f_2, \cdots, f_k\}
\]  

(2)

Each object \(O\) is defined as a vector of features \(F\) plus a class indicator \(C\) which we consider without loss of generality to be either 0 or 1. Thus, \(O = (C, F)\) is how we show objects and \(M : P(O) = P(C, F)\) is an example of a model. We also define the vector \(\theta\) as the set of parameters of the model \(M\) that produce the probability of observing a sequence of values for a specific dependent subgroup. The model class of \(k\) features \(M\), is the set of all possible models that partition the feature vector \(F\).

4 Probability calculation for classification

Having observed a sequence of \(n\) objects \(O^n\), our goal is to find the model that can produce this sequence with the highest probability compared to the other possible models.

\[
\arg \max_{M \in \mathfrak{M}} P(O^n|M) = \arg \max_{M \in \mathfrak{M}} P(C^n, F^n|M)
\]  

(3)

Using the Bayes rule, we can write (3) as

\[
\arg \max_{M \in \mathfrak{M}} P(C^n|M) P(F^n|C^n, M)
\]  

(4)

Since all the models in \(\mathfrak{M}\) produce the same probability for the sequence of classes \(C^n\), the first term in (4) does not effect the procedure of finding the maximum. Then the equation simplifies to

\[
\arg \max_{M \in \mathfrak{M}} P(F^n|C^n, M)
\]  

(5)

The set of partitioning subgroups \(S\) of model \(M\) defines the way \(P(F^n|C^n, M)\) is calculated. Namely

\[
P(F^n|C^n, M) = \prod_{i=1}^{g} P(F^n_{s_i}|C^n)
\]  

(6)

Where if for example \(s_i = \{f_3, f_4, f_5\}\) then

\[
P(F^n_{s_i}|C^n) = P(f^n_3, f^n_4, f^n_5|C)
\]  

(7)

To calculate the above probability for a specific model we need the parameter vector \(\theta\) of that model. For each of the models in the set \(\mathfrak{M}\) this vector is unknown to us. Lets consider the example of Naïve Bayes model. According to equation (5), we would like to calculate

\[
P(F^n|C^n, M_{Naïve}) = P(F^n_{s_1}, F^n_{s_2}, \cdots, F^n_{s_k}|C^n, M_{Naïve})
\]  

(8)

which is equal to

\[
P(F^n_{s_1}, F^n_{s_2}, \cdots, F^n_{s_k}|C^n, M_{Naïve}) = \\
P(f^n_1|C^n, M_{Naïve}) P(f^n_2|C^n, M_{Naïve}) \cdots P(f^n_k|C^n, M_{Naïve})
\]  

(9)

For a simple example we consider binary features. For each of the terms in the right hand side of the above equation, we need a parameter that determines the probability of
0 or 1 for that feature. This parameter would be an element of the vector $\theta$ of the Naïve model. For the first term on the right hand side we can write

$$P(f_1^n | C^n, M_{Naïve}) = \int_{\theta_1} P(f_1^n | \theta_1, C^n, M_{Naïve}) P(\theta_1 | C^n, M_{Naïve}) \, d\theta_1 \quad (10)$$

Not knowing the parameter $\theta_1$, to estimate probability of equation (10), we can use Krichevsky-Trofimov(KT)-estimator [6] due to the independence of the objects. This estimator is optimal because of the fact that it can asymptotically minimize the regret in the worst case of parameter $\theta$. This means that $P(\theta_1 | C^n, M)$ the prior over $\theta_1$ will be weighted over all data with a $(1/2, 1/2)$-Dirichlet distribution. Equation (11) shows the KT-estimator’s equation

$$P_e(f^n) = \int_{0}^{1} \frac{1}{\pi \sqrt{\theta(1-\theta)}} \theta^{n_1} (1-\theta)^{n_0} \, d\theta \quad (11)$$

where $n_1$ and $n_0$ are respectively the number of ones and zeros in the sequence.

Now that we can calculate the probability of observing a sequence of each dependent subgroup, by multiplying dependent subgroup probabilities of each model we have the probability of the sequence being produced according to that model. In the next section, we will introduce an efficient method to find the model that has the maximum probability of producing the observed sequence.

5 Network method and Maximum Likelihood model

The Network method is a way to compare different combinations of feature dependencies in a way that is efficient in terms of operations. The upmost row of nodes in the network represent the probability of observing the sequence of each element of the feature vector independently. For the sake of simplicity, we show $P_e(f^n_x, f^n_y)$ by $P_{xy}$. The nodes in the second row compare the sequence probability of the features shown in that node assuming them as dependent, with the multiplication of probabilities from the upper nodes which means independency between those features. The comparisons continue in the same procedure to the last node. Probabilities are calculated using KT estimation. Figure 2 shows a sample network graph.

![Network of 4 features](image)

Fig. 2. Network of 4 features

What follows explains how the maximum likelihood model emerges in the lowest node.
The maximal probability in each node is shown by $N$.

$$N_{1234} = \max\{P_{1234}, N_{123}, N_{124}, N_{134}, N_{234}\}$$

$$= \max\{P_{1234}, \max(P_{123}, N_{1}N_{23}, N_{1}N_{3}) \cdot P_{4}, \max(P_{12}, P_{12}P_{2}) \cdot \max(P_{34}, P_{3}P_{4}),$$

$$P_{1} \cdot \max(P_{234}, N_{2}N_{34}, N_{2}N_{4})\}$$

$$= \max\{P_{1234}, P_{123}P_{4}, P_{1} \cdot \max(P_{23}, P_{2}P_{3}) \cdot P_{4}, \max(P_{12}, P_{1}P_{2}) \cdot P_{3}P_{4},$$

$$P_{12}P_{34}, P_{12}P_{3}P_{4}, P_{1}P_{2}P_{3}P_{4}, P_{1}P_{2}P_{3}P_{4}, P_{12}P_{34}, P_{1}P_{2}P_{3}P_{4}, P_{1}P_{2}P_{3}P_{4}, P_{1}P_{2}P_{3}P_{4}\}$$

$$(12)$$

Accordingly, a node processes a subset of features from the feature vector. The probability assigned to that node is the largest out of the probabilities of all possible partitioning subgroups of those features. After calculating the node probabilities with this method, in the lowest node appears the ML probability model which is the result of a comparison out of all model probabilities in $M$.

In order to prove this, we give an example. Let’s consider the case that the ML model has the set of partitioning subgroups $S = \{s_{1} = \{f_{1}\}, s_{2} = \{f_{2}\}, s_{3} = \{f_{3}, f_{4}\}\}$ and in the last node of the graph, $P_{12}P_{34}$ is formed. This calculates the probability for a model that is composed of two partitioning subgroups $s_{1} = \{f_{1}, f_{2}\}$ and $s_{2} = \{f_{3}, f_{4}\}$. Figure 3 depicts this example with the irrelevant routes and nodes removed.

![Fig. 3. Selected model route in an example Network graph](image)

We can track the route that this model probability passes until it is formed in the lowest node. Out of the three pairs of arrows that end in the lowest node, we see that $P_{12}P_{34}$ is formed by the second pair. Tracking it up the graph, on the right arrow of that pair we reach the node $N_{34}$. In this node, $P_{34}$ and $P_{3}P_{4}$ are compared and $P_{34}$ is selected as the ML probability because of being larger. Following the left arrow of the same pair, we reach the node $N_{12}$. In this node, $P_{12}$ is compared to $P_{1}P_{2}$. Because we have assumed $S = \{s_{1} = \{f_{1}\}, s_{2} = \{f_{2}\}, s_{3} = \{f_{3}, f_{4}\}\}$ as the set of partitioning subgroups of the ML model, the probability $P_{1}P_{2}$ is larger than $P_{12}$. But since $P_{12}P_{34}$ is formed in the lowest node of the graph, $P_{12}$ must have been selected as ML probability in this node. This is in contradiction with the fact that the network method selects the larger probability in each node. Therefore it is impossible that $P_{12}P_{34}$ is formed in the lowest node of the graph instead of $P_{1}P_{2}P_{34}$. This proof can be repeated with any other arbitrary model being formed in the lowest node instead of the ML model. Thus, the constitution of the ML model probability in the lowest node is proved by contradiction.
Table I compares the amount of operations required by the network method to find ML model with those of direct computation of ML model for k features.

### Table I
**REQUIRED NUMBER OF OPERATIONS COMPARISON**

<table>
<thead>
<tr>
<th></th>
<th>Network method</th>
<th>Direct computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiplications</td>
<td>((k-1)k(k+1))</td>
<td>((k - 1)2^{k-2})</td>
</tr>
<tr>
<td>Comparisons</td>
<td>(\frac{(k-1)k(k+1)}{6})</td>
<td>(2^{k-1} - 1)</td>
</tr>
</tbody>
</table>

In Figure 4 the number of operations is plotted as a function of the number of features.

![Fig. 4. Computation workload comparison](image)

**6 Experiment and Results**

To compare the performance of the network method with the Naïve Bayes method, we generate sequences of feature vectors. Each feature vector contains 7 features. The set of partitioning subgroups is of the form \(S = (s_1 = \{f_1\}, s_2 = \{f_2, f_3\}, s_3 = \{f_4\}, s_4 = \{f_5, f_6, f_7\})\). Parameters \(\theta\) of these feature vectors are set manually in the simulation.

![Fig. 5. Probability of producing a sequence, using Network method compared to Naïve Bayes method](image)
In figure 5, the Maximum Likelihood probability of producing the sequence as a function of the length of the sequence is plotted. The probability graphs of producing each sequence using Naïve Bayes method and Network method are also plotted. The graphs show that the network method produces the sequence with a higher probability than the Naïve Bayes method. As the length of the sequence increases, the performance of Naïve Bayes method degrades while the Network method probability is quite close to ML estimation. The difference between the Network method and the ML estimation probability graphs is because of the unknown \( \theta \) parameters in the Network method. There is no way of knowing these parameters unless the sequence is produced synthetically.

When we have a finite set of models, the Bayesian Information Criterion (BIC) is a likelihood based criterion that can help us in the process of model selection. By introducing a penalty term proportional to the number of parameters in the model, BIC prevents an increase of likelihood that leads to over-fitting. Equation (13) shows how BIC is calculated for a model \( M \)

\[
BIC_M = -2 \times LLF_M + |\theta_M| \times \ln(n)
\]

where \( LLF_M \) is the log-likelihood function for model \( M \), \( |\theta_M| \) is the size of the parameter vector \( \theta \) of the model and \( n \) is the number of objects. On further consideration, this equation shows that a model with a smaller BIC has a better compromise between likelihood and complexity. In the case of the Network method, BIC proves that the model selected is not necessarily the most complex model possible. In the model class \( \mathfrak{M} \), the model that assumes every feature to be dependent has the largest parameter vector \( \theta \). Therefore it is considered the most complex model in the class and will have a larger penalty term when we use the BIC. Figure 6 compares the BIC for Naïve Bayes model, Network model and the model in which all the features are dependent.

As shown in figure 6, for shorter length of sequence, a simple model like Naïve Bayes model can fit the data with a low BIC. But as the length of the sequence increases a more complex model produces a higher likelihood and outperform Naïve Bayes model. For long enough sequences, the model selected with the Network method has the lowest BIC.
7 Conclusion

We have presented a method to find the Maximum Likelihood model among a large but finite set of models with less algebraic operations than the direct calculation. The selected model is the least complex one that can reproduce the observed sequence of data while capturing dependencies between the features. Therefore, when used in feature-based classification applications, will result in higher classification accuracy and lower workload. By using the Bayesian Information Criterion, we have proven that the model selected with our method is not always the one that has the largest number of parameters nor the simplest one but the one that fits the data best, independent of the data complexity. The method introduced in this paper leaves some intriguing matters to explore later. For example, can we apply restrictions to the complexity of the model selected? In other words, how can we limit the size of the set of partitioning subgroups? The answer to this question might require truncating the network after some row of nodes or eliminating some of the nodes in the lower rows. Another question would be whether it is possible to add weighting factors to the routes of the network, so that the selected model in the last node is the maximum a-posteriori model.

REFERENCES

Evaluation of Instrumental Measures for the Prediction of Musical Noise in Enhanced Noisy Speech

M. R. J. Gerrits¹, R. C. Hendriks¹, N. D. Gaubitch¹, J. Jensen²,³, and M. S. Pedersen²

¹ Delft University of Technology, 2628 CD, Delft, The Netherlands
² Oticon A/S, 2765 Sørum, Denmark
³ Aalborg University, 9100 Aalborg, Denmark

Abstract. Instrumental measures have been primarily used in order to predict speech quality or speech intelligibility. The aim of the present work is to evaluate the performance of a broad range of established instrumental measures in terms of their ability to predict the amount of musical noise present in enhanced noisy speech signals. The considered instrumental measures were evaluated using musical noise quantity scores obtained from a specially designed listening-test which was performed by normal-hearing listeners. Of all considered instrumental measures, a mean squared distortion measure, the PESQ measure, and the STOI measure yielded the highest correlations. These results confirm the ability of instrumental measures to predict the amount of musical noise, but further evaluation shows limitations to their applicability as musical noise predictors. However, subsequently the results suggest that simultaneous optimization for the amount of musical noise and speech quality or speech intelligibility is not possible using a single instrumental measure.

1 Introduction

The interest in the prediction of speech quality and speech intelligibility has led to the development of multiple instrumental speech-quality and speech-intelligibility measures. This development is supported by the demand for replacement of time consuming and expensive listening tests. However, as instrumental measures are not perfect, both instrumental measures as well as listening tests, are essential to judge the performance of speech enhancement algorithms. Over the years, many noise suppression algorithms have been introduced attempting to find optimal noise suppression without introducing loss of speech-quality or intelligibility under a diversity of distortion measures, see e.g., [1,2]. The performance of such speech enhancement algorithms is in general a trade-off between reduction of the noise and the extent in which the algorithm distorts the speech signal. Finding the right tradeoff which is crucial in the field of hearing aids, where the sound to be presented to the hearing impaired listener should be as natural and with as high quality as possible.

However, removing all the noise in a noisy speech signal without introducing loss of speech quality requires that the noise is known exactly. This is unrealistic for any practical application. Therefore, many noise reduction methods employ the statistics of the speech and noise processes, e.g. the speech and noise power spectral densities. As a result of working with statistical descriptors of the signal, instead of the actual realizations, artifacts are introduced during the noise reduction process. Among these is the highly annoying residual noise known as musical noise [3], which can remain after processing, and decrease the quality and speech intelligibility of the enhanced speech signal. Although it is possible to hide or reduce the musical noise to some extent by adjusting certain parametric settings of the noise reduction algorithm [3,4], this is not straightforward without a clear instrumental measure.

While some instrumental measures have shown to predict the quality of enhanced noisy speech with high correlation (see [5], [6], [7]), only little is known about the ability of instrumental measures to predict the amount of musical noise, even though this is an important aspect of speech quality. The aim of this contribution is to evaluate the ability of an instrumental measures to predict the amount of musical noise in the enhanced signal. Given that there are measures which are able to predict the amount of musical noise, these could be used to draw additional conclusions about the performance of noise reduction methods and be of use in their development.

In 2008, Uemura et. al [8] proposed a novel measure based on the kurtosis of the power spectral density (PSD) of the enhanced noisy speech signal which can be used specifically for predicting the amount of musical noise in the enhanced speech signal. Besides this measure, we consider various other established instrumental measures in terms of their ability to predict the results of a specially designed
listening experiment. The results of the subjective listening-test characterize the quantity of musical noise in several enhanced noisy speech signals generated by applying the spectral subtraction noise reduction method (SS), and varying the amount of noise reduction. Evaluation of the prediction performance of the different instrumental measures is based on Pearson’s correlation coefficient and the Kendall rank correlation coefficient [9].

2 On spectral subtraction and musical noise production

We consider a noisy speech model that can be described as a clean speech signal $s(n)$ that is degraded by an additive uncorrelated noise signal $v(n)$, i.e., $y(n) = s(n) + v(n)$, where $y(n)$ represents the observed noisy speech with time-sample index $n$. It is assumed that both $s(n)$ and $v(n)$ are wide-sense stationary for short time segments, and a DFT based analysis-modification-synthesis procedure is applied, with the aim of reducing noise. After applying a short-time Fourier transform (STFT) to the noisy signal, the noisy observation can be described by,

$$Y(i, k) = S(i, k) + V(i, k),$$

where upper case letters describe the signal in frequency domain. The frame-index and transform coefficients are described by $i$ and $k$ respectively. The spectral subtraction noise reduction method [3,10] can be used to modify the frequency spectrum of the noisy observation. By subtracting a noise PSD estimate, $\sigma^2_Y(i, k)$, from a noisy observation, $|Y(i, k)|^2$, a clean speech periodogram, $|\hat{S}(i, k)|^2$, is estimated. The SS noise reduction procedure is described by,

$$|\hat{S}(i, k)|^2 = \begin{cases} H(i, k)|Y(i, k)|^2 & \text{for } H(i, k)|Y(i, k)|^2 > \alpha_{ss}\sigma^2_Y(i, k) \\ \alpha_{ss}\sigma^2_Y(i, k) & \text{otherwise} \end{cases}$$

where $H(i, k)$ is given by,

$$H(i, k) = 1 - \frac{\beta_{ss}\sigma_Y^2(i, k)}{|Y(i, k)|^2}$$

$\beta_{ss}$ is called the over-subtraction factor and represents a scalar weight on $\sigma_Y^2(i, k)$. $\alpha_{ss}$ denotes a flooring parameter which introduces a minimum value for gain, $H(i, k)$. After modification the signal is reconstructed using an inverse STFT.

However, even when the noise PSD is perfectly known, it is impossible to remove all the noise as this requires all noise realizations to be known. These variations of the power of these noise realizations around its PSD will introduce estimation errors in $H(i, k)$. These estimation errors result in variations in $H(i, k)$ with musical noise as a consequence. Consider a noise-only segment, $y(n) = v(n)$, where $|V(i, k)|^2 > \beta_{ss}\sigma_Y^2(i, k)$ thus $H(i, k) > 0$, this means that part of the noise power will remain in the enhanced periodogram. If this power residue is isolated in both time and frequency, i.e. a narrow-band spectral peak, it will introduce a short tonal artifact after reconstruction of the signal. Whenever isolated spectral peaks occur in multiple time-frames across different frequency bands, the residual noise is described as musical noise. Generally, altering $\beta_{ss}$ and $\alpha_{ss}$ provides a trade-off between the amount of musical noise, broadband noise reduction and speech distortion.

3 Instrumental measures

A broad range of established instrumental measures is used with the objective to evaluate their ability to predict the amount of musical noise in enhanced noisy speech. Included are the often used Perceptual Evaluation of Speech Quality measure, the log-likelihood ratio, the segmental SNR, the frequency weighted segmental SNR, and the normalized frequency weighted segmental SNR. But also several spectral distance measures are taken into account, i.e. the weighted spectral slope, the cepstral distance, the log-spectral distance, two mean squared error measures, the Itakura-Saito distance, a COSH distance measure based on the IS measure, and two composite measure for noise distortion, which is based on a combination of seven objective SQ measures. Additionally we apply one musical noise measure, based
on the kurtosis-ratio between the PSD of the enhanced and the noisy speech signal, and one speech-intelligibility measure, the short-time objective intelligibility measure. Table (1) provides a summary of all instrumental measures with their corresponding abbreviations. For the PESQ, LLR, fwSEG, fwSEGn, IS1, CEP, WSS and COMP measure we apply the implementations as provided by Loizou in [2], whereas for IS2, segSNR and the COSH metric we apply the implementation as provided in the VOICEBOX toolbox [11]. The IS1 and IS2 measures are the same, however IS1 is an LPC based implementation, where IS2 is calculated using the PSD of the clean and enhanced speech. Both these IS measures and the COSH measure have been limited to 100. The two different Mean Squared Distance methods, $d_{MSD1}$ and $d_{MSD2}$, are implemented based on respectively the complex DFT coefficients ($X = S$), and the periodogram ($X = |S|^2$) of the estimated clean and noisy speech signal.

$$
d_{MSD}(X) = \frac{1}{M} \sum_{k} \sqrt{\frac{1}{L} \sum_{i} |X(i,k) - \hat{X}(i,k)|^2}
$$

where $M$ describes the total number of frequency coefficients, and $L$ denotes the total number of time-frames. The ‘kurtosis ratio’ measure is described by the ratio between the kurtosis of the PSD of the noisy signal, $\text{kurt}_{\text{noisy}}$, and the kurtosis of the PSD of the enhanced signal, $\text{kurt}_{\text{enhanced}}$, i.e.,

$$
d_{\text{kurtosis-ratio}} = \frac{\frac{1}{M} \sum_{k} \text{kurt}_{\text{enhanced}}(k)}{\frac{1}{M} \sum_{k} \text{kurt}_{\text{noisy}}(k)},
$$

where $\text{kurt}$ is defined by,

$$
\text{kurt}(k) = \frac{1}{(\frac{7}{3})^4} \sum_{i} X(i,k)^4
$$

Dependent on whether $\text{kurt}_{\text{noisy}}$ or $\text{kurt}_{\text{enhanced}}$ is computed, $X$ represents either the realizations of the noisy signal, $|Y(k)|^2$, or the realizations of the enhanced speech signal, $|\hat{S}(k)|$. In [8] it was stated that a high kurtosis-ratio corresponds to a high amount of musical noise, i.e. the distribution of the isolated residual noise coefficients correspond to a high kurtosis values, which indicates a distribution

<table>
<thead>
<tr>
<th>Instrumental Measures</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PESQ</td>
<td>Perceptual Evaluation of Speech Quality [12]</td>
</tr>
<tr>
<td>LLR</td>
<td>Log-Likelihood ratio [13]</td>
</tr>
<tr>
<td>segSNR</td>
<td>Segmental SNR [13]</td>
</tr>
<tr>
<td>fwSNR</td>
<td>Frequency weighted segmental SNR [13]</td>
</tr>
<tr>
<td>fwSNRn</td>
<td>normalized Frequency weighted segmental SNR [5]</td>
</tr>
<tr>
<td>LSD</td>
<td>Log-Spectral Distance [14]</td>
</tr>
<tr>
<td>MSD1</td>
<td>Mean Squared Distance</td>
</tr>
<tr>
<td>MSD2</td>
<td>Mean Squared Distance (PSD)</td>
</tr>
<tr>
<td>IS</td>
<td>Itakura-Saito Distance [13]</td>
</tr>
<tr>
<td>COSH</td>
<td>Symmetric version of IS</td>
</tr>
<tr>
<td>CEP</td>
<td>Cepstral Distance [13]</td>
</tr>
<tr>
<td>WSS</td>
<td>Weighted Spectral Slope [13]</td>
</tr>
<tr>
<td>STOI</td>
<td>Short-Time Objective Intelligibility Measure [14]</td>
</tr>
<tr>
<td>KURT</td>
<td>Kurtosis ratio [8]</td>
</tr>
<tr>
<td>COMPOv1</td>
<td>Composite Noise Distortion Measure, overall quality [5]</td>
</tr>
<tr>
<td>COMPbn</td>
<td>Composite Noise Distortion Measure, background distortion [5]</td>
</tr>
</tbody>
</table>

Table 1. The various different instrumental measures used for evaluation
with a sharp peak. However, this measure contains some practical limitations, e.g. the proposed measure is solely applied on noise-only segments, [8]. This is solved using a binary mask based on a SNR threshold of 0 dB. Also note that calculating the frequency average in Eq. (4,6), does only make sense for distributions which are spectrally flat.

4 Listening experiments

The data-set, Q, used for this experiment includes a set of 20 different clean speech estimates, $\tilde{s}(n)$. These clean speech estimates are generated by applying the SS noise reduction method, Eq. (2), to a noisy speech signal, and varying the subtraction-factor $\beta_{ss}$. Hence, these enhanced signals contain different amounts of musical noise. As masking of residual background noise is undesired for this specific experiment, the spectral floor, $\alpha_{ss}$, is set to 0.

The clean speech, $s(n)$, used for generating the stimuli, consisted of a concatenation of two sentences which originated from the TIMIT [15] database sampled at 16 kHz. The two different sentences where respectively read by a female and a male, with a combined total signal length of approximately 7 seconds.

A noisy speech signal, $y(n)$, is produced by degrading this clean speech signal with additive white Gaussian noise, $v(n)$, and the signal-to-noise ratio is set to 5 dB. The noise PSD is estimated over 10 seconds of noise-only signal. All noise reduction is performed in the power spectral domain, using time frames of 32 ms with 50% overlap and a 512 point FFT is used to transform each time-frame into the DFT domain, and a square-root Hann window is used as analysis and synthesis window. As stated before, altering the over-subtraction factor, $\beta_{ss}$, allows changing the amount of musical noise. The over-subtraction factor is varied in steps of 0.5 from $\beta_{ss} = 0$, i.e. no noise reduction, to $\beta_{ss} = 9.5$ where the estimated clean speech consists of highly distorted speech and little to no musical noise due to the large over-subtraction. Such a large value of $\beta_{ss}$ is rarely in practice due to the amount of speech distortion that it introduces. However, it is an important for a musical noise predictor to be able to identify situations in which little or no musical noise remains.

The listening test is designed to characterize the amount of musical noise in an enhanced noisy speech signal. Listeners were asked to specifically grade the amount of musical noise, by answering the following two, equivalent, questions: 'How much musical noise is present in the stimuli? That is, How musical is the background noise perceived?'. However, comparison of the 20 enhanced signals at once is too tiring for the listener, therefore, the data-set, Q, is split into two different subsets of each 10 enhanced speech signals, $Q_{set1}$ and $Q_{set2}$, respectively containing the enhanced signals, $\tilde{s}(n)$ generated with over-subtraction factors $\beta_{ss} \in \{0, 1, 2, ..., 9\}$, and $\beta_{ss} \in \{0.5, 1.5, 2.5, ..., 9.5\}$.

All subjects performed a training round, which consists of subset $Q_{set2}$ and two test rounds, where $Q_{set1}$ and $Q_{set2}$ are successively presented to the listener. The signals in each subset are presented simultaneously, compared among each other, and graded individually by assigning a single number in the range 1 to 5 with decimal steps of 0.1, providing a numerical indication of the amount of musical noise. That is, the listeners rate the amount of narrow-band residual noise present in the background of the presented stimuli by comparing the 10 presented signals. Grade 1 is assigned to signals with no musical noise e.g. broadband noise or no background distortion, while grade 5 is assigned to signals with extreme amounts of musical noise. The ratings are depicted in Table 2.

<table>
<thead>
<tr>
<th>score</th>
<th>The amount of musical noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Extreme</td>
</tr>
<tr>
<td>4</td>
<td>A lot</td>
</tr>
<tr>
<td>3</td>
<td>Medium</td>
</tr>
<tr>
<td>2</td>
<td>A little</td>
</tr>
<tr>
<td>1</td>
<td>Broadband or no musical noise</td>
</tr>
</tbody>
</table>

Table 2. Grading scale for the amount of musical noise
Prior to the actual listening test, 3 signals are presented to the listener to make the subject accustomed to the signals presented in the listening test. Two of these signals contain enhanced noisy speech with some degree of musical noise while the other signal contains the unprocessed noisy speech. Subjects are asked to explain the task at hand, to ensure complete understanding of the test. The listening test was performed in total by 20 subjects inside a sound proof listening room. The group of subjects consists 90% male, and 10% female listeners, within the age-range of 23 up to 51 years old. The average age of a listener was 30 years old, and the listeners were not hearing impaired to the best of their knowledge. Finally a mean opinion musical noise quantity score is calculated by averaging over the obtained scores. The average scores of the listening tests are depicted Figure 1.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Listening-test results for quantifying musical noise, i.e. The mean-opinion musical noise quantity score.}
\end{figure}

5 Evaluation

All instrumental measures (Table 1) are evaluated using data-set $Q$, after which we apply two different performance metrics to measure the performance in terms of their ability to predict the amount of musical noise. Performance measures are commonly used for evaluating correlations between instrumental and listening test data, see [2]. First, the Pearson correlation coefficient, $\rho$, is included,

\[ \rho = \frac{\sum_l (S_l - \bar{S})(D_l - \bar{D})}{\sqrt{\sum_l (S_l - \bar{S})^2 \sum_l (D_l - \bar{D})^2}}. \]  

Here, $S$ and $D$ describe, respectively, the mean opinion musical noise quantity scores (Figure 1) and the instrumental scores. $\bar{S}$ and $\bar{D}$ define the means of the sets $S$ and $D$, and $l$ denotes the over-subtraction index. Secondly we imply Kendall’s tau rank correlation coefficient,

\[ \tau = \frac{N_c - N_d}{\frac{1}{2}N(N-1)}. \]  

Here, $N_c$ describes the concordant pairs, while $N_d$ describes the discordant pairs [9]. $N$ denotes the total number of enhanced speech signals applied for the subjective test, i.e. $N = 20$. This latter measure is used to support the conclusions made based on the correlation coefficient $\rho$.

To compare the performance of the different instrumental measures, the absolute value of both coefficients is applied, i.e. $|\rho|$ and $|\tau|$ must be in the range of $0 < |\rho| < 1$ or $0 < |\tau| < 1$. Here, a coefficient close to one corresponds to a high agreement between the instrumental measure and the subjective data, whereas a coefficient close to zero indicates the data to be independent. An overview of the performance of all instrumental measures is shown in Figure 2.
6 Results and Discussion

The results displayed in Figure 2 indicate that the STOI, PESQ and MSD2 measure provide a good prediction of the amount of musical noise present in the enhanced signal, there a strong correlation exists between the output of these instrumental measures and the listening test scores. Therefore, we will evaluate the performance of these three methods in more detail. The results of these three metrics are depicted in Figure 3, illustrated by the dashed line. To gain insight into how the gain function, $H(i, k)$, distorts the clean speech, we apply gain $H(i, k)$ to the clean speech $|S(i, k)|^2$, and calculate the results for the three instrumental metrics. These results are portrayed in Figure 3, and illustrated by the dotted line. And immediately a rather important observation can be made when comparing the dashed and dotted curves for each of the three measures. Although the dashed curves of the PESQ and the STOI measure provide a strong correlation with the listening test results, we can not conclude that these two metrics are good measures for predicting musical noise in general. Note that the dotted curves represent processed clean speech, which does not contain any musical noise. However, for these signals the PESQ and the STOI measure produce high results, and therefore it is not possible to predict the amount of musical noise without prior knowledge on the presence of musical noise. This limits the applicability of the PESQ and the STOI measure as a musical noise predictor drastically. Considering the performance of MSD2, it can be seen that this measure is sensitive to scaling of the enhanced noisy speech signal, there scaled versions of the enhanced noisy speech signal will produce different outputs. However, the perception of the sounds will not change.

Consider the minimum of the MSD2 measure as a function of $\beta_{ss}$, here the average distance between $|S(i, k)|^2$ and $|\hat{S}(i, k)|^2$ is small, i.e. the average distortion is small. But note, comparing Figure 1 and 3, that whenever the MSD2 at its minimum, the amount of musical noise is high. This indicates that the best clean speech estimate contains according to this instrumental measure the largest amount of musical noise. A similar conclusion can be drawn when evaluating the results of the PESQ measure, there the
maximum mean opinion scores, i.e. the maximum speech quality scores, are predicted for $\beta_{SS}$ values that lead to high amounts of musical noise. Likewise, in case of the STOI measure, the maximum speech intelligibility scores are predicted whenever the amount of musical noise is high according to Figure 1. This means that when optimizing the enhanced speech signal for intelligibility, quality or the distortion, the signal will contain a high amount of musical noise. Subsequently, optimizing for both the minimum amount of musical noise and the best speech quality simultaneously is not possible using the instrumental measures considered in this paper.

Another observation can be made, based on the results of the PESQ and STOI measure. From Figure 3 it can be seen that it is possible to apply high values of $\beta_{ss}$ and largely reduce the amount of musical noise, without losing a lot of speech quality, and on account of introducing only a limited decrease in speech intelligibly.

Additionally we can conclude from Figure 2 that the background distortion measure, COMPbn, and the musical noise prediction method, KURT, do not correlate well with the listening test results. In case of the COMPbn measure this result is consistent with the observations made in [5]. However, in case of the KURT measure this is a more unexpected result, but can be explained as follows. First consider Equation (6). If $\beta_{ss}$ is increased, then more and more spectral values will be set to zero, and the PDF of the enhanced speech will become increasingly peaky. As a consequence, the kurtosis of the PDF of the enhanced signal will increase until all spectral values are set to zero, for which the method becomes undefined. By comparing this with the results in Figure 1, it can be observed that the increasing $d_{kurt-ratio}$ has a strong correlation for the range $0 < \beta_{ss} < 2$, which is consistent with the experiments in [8]. However, for the range $\beta_{ss} > 2$ the $d_{kurt-ratio}$ metric keeps increasing, where the listening test results show a decrease in the amount of musical noise. We can thus conclude that the KURT metric performs poorly for values of $\beta_{ss} > 2$.

Form Figure 2 we observe that a large group of instrumental methods show low correlation with the listening test results. This could be caused by the fact that both $\tau$ and $\rho$ only provide a metric to illustrate a linear relation. In order to obtain a better correlation between the results of the instrumental measures and the listening test scores, a mapping could be used to account for nonlinear relations, see [14].

7 Conclusions

The results provided in this paper concluded that it is possible to predict the amount of musical noise in an enhanced noisy speech signal by either using a mean squared distance measure, MSD2, the PESQ measure or the STOI measure. There the results of these instrumental methods indicate strong correlations with the listening test results. However, this high correlation suggests that simultaneous optimization for the minimum amount of musical noise and maximum speech quality or speech intelligibly is not possible using a one of the considered instrumental measure. Surprisingly, the instrumental method, KURT (Table 1), which was specially designed for predicting the amount of musical noise shows low correlation with the listening test results. On basis of the analysis that existing instrumental methods are limited in the prediction of musical noise, we can speculate on the need for a better metric specifically designed to measure the musical noise quantity.

References

<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahmad U.</td>
<td>Jansen C.J.A.</td>
<td>Ren Z.</td>
</tr>
<tr>
<td>Alpar G.</td>
<td>Janssen E.J.G.</td>
<td>Rijmen V.</td>
</tr>
<tr>
<td>Baggen C.P.M.J.</td>
<td>Jensen J.</td>
<td>Rodriguez J.</td>
</tr>
<tr>
<td>Baltus P.G.M.</td>
<td>Jin Z.</td>
<td></td>
</tr>
<tr>
<td>Bauduin M.</td>
<td>Karsmakers P.</td>
<td>Schreurs D.</td>
</tr>
<tr>
<td>Bergmans J.W.M.</td>
<td>Kononchuk D.</td>
<td>Singelee D.</td>
</tr>
<tr>
<td>Bilgin B.</td>
<td>Lagendijk R.L</td>
<td>Skoric B.</td>
</tr>
<tr>
<td>Bodart J.</td>
<td>Lampe L.</td>
<td>Soh P.J.</td>
</tr>
<tr>
<td>Bourdoux A.</td>
<td>Lauwereins r.</td>
<td>Spreeuwers L.J.</td>
</tr>
<tr>
<td>Chiumento A.</td>
<td>Lazov V.</td>
<td>Stokking H.</td>
</tr>
<tr>
<td>Deleu T.</td>
<td>Leroux P.</td>
<td></td>
</tr>
<tr>
<td>De Doncker P</td>
<td>Li C.</td>
<td></td>
</tr>
<tr>
<td>de Groot J.</td>
<td>Li M.</td>
<td></td>
</tr>
<tr>
<td>De Meyer L.</td>
<td>Linnartz J.-P.</td>
<td></td>
</tr>
<tr>
<td>Desset C.</td>
<td>Louveaux J.</td>
<td></td>
</tr>
<tr>
<td>Determe J.-F.</td>
<td>Luks W.</td>
<td></td>
</tr>
<tr>
<td>Dicot J.M.</td>
<td>Luz P.</td>
<td></td>
</tr>
<tr>
<td>Drijvers M.</td>
<td>Marin E.</td>
<td></td>
</tr>
<tr>
<td>Duport F.</td>
<td>Martin A.</td>
<td></td>
</tr>
<tr>
<td>Duran R.</td>
<td>Massar S.</td>
<td></td>
</tr>
<tr>
<td>Erkin Z.</td>
<td>Mercuri M.</td>
<td></td>
</tr>
<tr>
<td>Gastpar M.</td>
<td>Moonen M.</td>
<td></td>
</tr>
<tr>
<td>Gaubitch N.D.</td>
<td>Mosk A.P.</td>
<td></td>
</tr>
<tr>
<td>Gerrits M. RJ.</td>
<td>Nguyen D.</td>
<td></td>
</tr>
<tr>
<td>Gokberk B.</td>
<td>Ors S. Berna</td>
<td></td>
</tr>
<tr>
<td>Goseling J.</td>
<td>Ozkan M. Akif</td>
<td></td>
</tr>
<tr>
<td>Habibi H.</td>
<td>Papagiannopoulos K.</td>
<td></td>
</tr>
<tr>
<td>Hendriks R.C.</td>
<td>Pederssen M.S.</td>
<td></td>
</tr>
<tr>
<td>Hendriks A.</td>
<td>Peng Y.</td>
<td></td>
</tr>
<tr>
<td>Horlin F.</td>
<td>Pinkse P.W.H.</td>
<td></td>
</tr>
<tr>
<td>Hsin-Hung C.</td>
<td>Pollin S.</td>
<td></td>
</tr>
<tr>
<td>Jalarirad A.</td>
<td>Preneel B.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>