SIGNAL 2025

The Tenth International Conference on Advances in Signal, Image and Video Processing

ISBN: 978-1-68558-245-6

March 9th –13th, 2025

Lisbon, Portugal

SIGNAL 2025 Editors

Pavel Loskov, ZJU-UIUC, China

SIGNAL 2025

Foreword

The Tenth International Conference on Advances in Signal, Image and Video Processing (SIGNAL 2025), held between March 9 - 13, 2025, continued the inaugural event considering the challenges mentioned above. Having these motivations in mind, the goal of this conference was to bring together researchers and industry and form a forum for fruitful discussions, networking, and ideas.

Signal, video and image processing constitutes the basis of communications systems. With the proliferation of portable/implantable devices, embedded signal processing became widely used, despite that most of the common users are not aware of this issue. New signal, image and video processing algorithms and methods, in the context of a growing-wide range of domains (communications, medicine, finance, education, etc.) have been proposed, developed and deployed. Moreover, since the implementation platforms experience an exponential growth in terms of their performance, many signal processing techniques are reconsidered and adapted in the framework of new applications. Having these motivations in mind, the goal of this conference was to bring together researchers and industry and form a forum for fruitful discussions, networking, and ideas.

We take here the opportunity to warmly thank all the members of the SIGNAL 2025 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to SIGNAL 2025. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the SIGNAL 2025 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that SIGNAL 2025 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of signal processing.

We are convinced that the participants found the event useful and communications very open. We also hope that Nice provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city

SIGNAL 2025 Chairs:

SIGNAL 2025 Steering Committee
Wilfried Uhring, Université de Strasbourg, France
Jérôme Gilles, San Diego State University, USA
Constantin Paleologu, Polytechnic University of Bucharest, Romania
Sergey Y. Yurish, Excelera, S. L. | IFSA, Spain
Pavel Loskot, ZJU-UIUC Institute, China

SIGNAL 2025 Publicity Chairs
Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain

José Miguel Jiménez, Universitat Politecnica de Valencia, Spain

SIGNAL 2025

Committee

SIGNAL 2025 Steering Committee
Wilfried Uhring, Université de Strasbourg, France
Jérôme Gilles, San Diego State University, USA
Constantin Paleologu, Polytechnic University of Bucharest, Romania
Sergey Y. Yurish, Excelera, S. L. | IFSA, Spain
Pavel Loskot, ZJU-UIUC Institute, China

SIGNAL 2025 Publicity Chairs
Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
José Miguel Jiménez, Universitat Politecnica de Valencia, Spain

SIGNAL 2025 Technical Program Committee

Waleed H. Abdulla, The University of Auckland, New Zealand
Ahmed Al Hilli, Technical College of Najaf | Al-furat Al-Awsat Technical University, Iraq
Kiril Alexiev, Institute for Information and Communication Technologies -Bulgarian Academy of Sciences, Bulgaria
Djamila Aouada, SnT | University of Luxembourg, Luxembourg
Nadia Baaziz, Université du Québec en Outaouais, Canada
Junaid Baber, University of Balochistan, Pakistan
Vesh Raj Sharma Banjade, Intel Coporation, USA
Joan Bas, CTTC, Spain
Wassim Ben Chikha, Tunisia Polytechnic School, Tunisia
Amel Benazza-Benyahia, SUP'COM | COSIM lab. | University of Carthage, Tunisia
Anirban Bhowmick, Vellore Institute of Technology | Bhopal University, India
Larbi Boubchir, LIASD - University of Paris 8, France
Moez Bouchouicha, LIS - Laboratoire d'Informatique et Systèmes | Toulon University, France
Salah Bourennane, Ecole Centrale de Marseille, France
Geraldo Braz, Federal University of Maranhão, Brazil
Valeria Bruschi, Università Politecnica delle Marche, Ancona, Italy
Paula María Castro Castro, University of A Coruña, Spain
M. Girish Chandra, TCS Research & Innovation, India
Zhuyun Chen, South China University of Technology,Guangzhou, China
Qiang Cheng, University of Kentucky, USA
Doru Florin Chiper, Technical University Gheorghe Asachi of Iasi, Romania
Sergio Cruces, Universidad de Sevilla, Spain
João Dallyson Sousa de Almeida, Federal University of Maranhão, São Luís, Brazil
Natasja M. S. de Groot, Erasmus Medical Center | Technical University Delft, Netherlands
Laura-Maria Dogariu, National University of Science and Technology POLITEHNICA Bucharest, Romania
António Dourado, University of Coimbra, Portugal

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

Combined EEG/ERG Features for Bipolar Disorders Diagnosis 1
Julie Muzzolon, Xiaoxi Ren, Steven Le Cam, Thomas Schwitzer, and Valerie Louis Dorr

A New 1D-CNN Paradigm for Onset Detection of Absence Seizures in Children 3
Maxime Yochum, Amar Kachenoura, Matthieu Aud'hui, Fabrice Wendling, Anna Kaminska, Rima Nabbout, Mathieu Kuchenbuch, and Pascal Benquet

An Integrative Strategy for Solving the EEG Inverse Problem and the Estimation of Brain Effective Connectivity in Epilepsy. A Proof-of-Concept Study. 8
Marc Greige, Ahmad Karfoul, Pascal Benquet, Maxime Yochum, and Regine Le Bouquin Jeannes

FMSTFnet: Feature-Modulation Spatio-Temporal Fusion Network for HDR Video 11
Wei Zhang, Yeyao Chen, and Gangyi Jiang

InterGridNet: An Electric Network Frequency Approach for Audio Source Location Classification Using Convolutional Neural Networks 16
Christos Korgialas, Ioannis Tsingalis, Georgios Tzolopoulos, and Constantine Kotropoulos

Camera Calibration and Stereo via a Single Image of a Spherical Mirror 22
Nissim Barzilay, Ofek Narinsky, and Michael Werman

Efficient Implementation of CNN in Deep Learning by Using Multirate Algorithms 28
Guowei Xiao, Yingshuai Wang, and Ping Wang

Finite Word Length Effect in Practical Block-Floating-Point FFT 33
Gil Naveh

Minecraft of System Modeling 40
Pavel Loskot

Combined EEG/ERG Features for Bipolar Disorders Diagnosis

Julie Muzzolon[1,4], Xiaoxi Ren[2], Steven Le Cam[1], Thomas Schwitzer[3], Valérie Louis Dorr[1]
1. Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France
2. Unité d'Imagerie Adaptative Diagnostique et Interventionnelle, Nancy, France
3. Pôle Hospitalo-Universitaire de Psychiatrie d'Adultes et d'Addictologie du Grand Nancy, Centre Psychothérapique de Nancy, Laxou, France
4. e-mail: julie.muzzolon@univ-lorraine.fr

***Abstract*—Bipolar Disorder (BD) is a disabling lifelong condition that remains misdiagnosed. Robust biomarkers are needed for a reliable and early diagnosis. Recent studies have demonstrated that electrophysiological ERG/EEG measurements hold relevant features for the diagnosis of BD. In this study, we propose a combined analysis of these modalities with promising performance for the detection of BD subjects with respect to controls.**

Keywords-ERG; EEG; DWT; bipolar disorders; SVM.

I. Introduction

Bipolar disorders (BD) are characterized by alternating manic and depressive episodes. Although these disorders are quite common, the diagnosis is often late [1] and subjective since it primarily relies on an interview guided by a clinician. Hence, there is a need for more robusts biomarkers independent of the subjective interpretations of patients and practitioners.

Previous studies have shown that psychiatric disorders in general affect the responses of retinal rod and cone cells [2]–[4], and that electroretinigram (ERG) responses to light stimuli can help in the differential diagnosis of mental disorders [5][6]. Electroencephalogram (EEG) alterations in responses recorded from primary visual cortex areas are also well-documented [7][8].The aim of this study is to assess the benefit of combining ERG with EEG measurements. To the best of our knowledge, no previous research work applied machine learning techniques to coupled ERG/EEG features for BD diagnosis.

Most studies focus on waveform amplitudes and latencies of a and b waves [5]. These temporal features are sensitive to noise and do not characterize the whole response waveforms. We then propose to extract time-frequency (TF) features from ERG and EEG responses using Discrete Wavelet Transform (DWT) [9]. The most significant coefficients according to the Wilcoxon rank sum test (alpha risk < 0.05) were selected.

Finally, we performed classification using Support Vector Machine (SVM) on 6 datasets : we studied the discriminating power of TF features against temporal features from EEG alone, ERG alone and combined ERG/EEG. Our database being rather modest in size, we performed stratified k-fold cross-validations to avoid overfitting. Averaged F1-score, Accuracy, Recall and Specificity scores are reported, as well as the standard deviation (SD) of these criteria over the tested folds.

In Section II, we introduce the data source and methods employed to collect the recordings, denoise the signals, extract the biomarkers and perform our predictions. In Section III, we describe the selected biomarkers and the prediction results. Finally, in Section IV, we conclude about the benefit of coupled ERG/EEG TF features in BD diagnosis.

II. Methods

A. Data source and protocol

ERG (right and left eyes averaged) and EEG (average of 4 electrodes over primary visual cortex of both hemispheres) responses to visual stimuli were recorded on euthymic bipolar patients (N = 30, Age (mean $\pm$ SD) = 47.5 ± 13.3, 67.7% women) and on healthy control subjects (N = 25, Age (mean $\pm$ SD) = 42.3 ± 14.8, 60.0% women) who were included in the BiMar study carried out by the CPN, Nancy, France. We used the Retinaute device (BioSerenity), a virtual reality headset fitted with electrodes that simultaneously records ERG and EEG responses. All stimuli were performed according to the International Society for Clinical Electrophysiology of Vision (ISCEV) standards [10][11].

We recorded ERG and EEG responses under dark-adapted (DA) and light-adapted (LA) conditions with a strength of 3.0 cd.s.m^{-2} (DA3.0, LA3.0). In total 16 and 32 flashes for DA3.0 and LA3.0 respectevily. A 30Hz flash LA3.0 (Flicker) was also repeated 16 times. Each stimulus triggers an electrical activity of a specific cell in the retina : the combined rod-cone activity can be studied with DA3.0 and cone activity only with LA3.0.

B. Signal denoising and preprocessing

50Hz powerline interference was removed with an infinite impulse response notch filter (center frequency = 50Hz, quality factor = 5). We did a 10-level DWT decomposition and set approximate coefficients and corresponding detail coefficients to zero to remove low frequencies (0-1 Hz) and high frequencies (above 62 Hz) [9]. The stimuli consisting of a repetition of flashes, we then segmented our signals into equal-size epochs starting 50 ms before each flash. Ouliers epochs were rejected and we worked on the averaged epoch.

C. Biomarkers selection

We selected the amplitude and latencie of a and b waves for DA3.0 and LA3.0 [11]. The retinal response to the Flicker stimulus is periodic, so we measure the amplitudes and latencies of the first trough and peak. The EEG responses result in a series of negative (N-waves) and positive waves (P-waves), but we focused on the P2-wave as it is the most robust [10].

In order to extract more relevant features, we computed a 6-level DWT analysis [9] that gives a synthetic and non

 ISBN: 978-1-68558-245-6

redundant representation of the ERG and EEG in both time and frequency domains. The sampling frequency of our signals being 1000 Hz, it allows us to analyze the energy content in the frequency ranges $[0, 8]$, $[16, 31]$, $[31, 62]$, $[62, 128]$, $[128, 256]$, and $[256, 512]$ Hz. We chose 'daubechies-4' wavelet since it gave the best reconstruction of our signals once the lowest energy coefficients were removed.

A nonparametric Wilcoxon rank sum test with an alpha risk of 0.05 was used to select coefficients significantly different between patients with BD and the healthy population.

D. Machine learning model and prediction evaluation

We conducted our classification on ERG, EEG and coupled ERG/EEG features. We analyzed wave time characteristics and TF coefficients separately. Classification was made using a linear SVM classifier that separates the two classes (1 = BD, 0 = controls)[12]. In order to evaluate the discriminating power of our model, we performed a stratified cross-validation, where our data set was randomly split into 5 folds within each the proportion of the classes is preserved : 4 folds constitute the training set (N = 44) and the 5^{th} fold is the test set (N = 11). We repeat this operation 10 times so we have 50 predictions for each dataset.

We recorded the accuracy, recall, specificity and F1-score at each step, then these scores are averaged. We also pay attention to variability in the predictions by computing the SD of the scores. A great recall (resp. specificity) means that only a few bipolar patients (resp. controls) will be misclassified.

III. Results

Temporal characteristics selection showed a significant greater a-wave amplitude for DA3.0 ($p < 0.05$) as long as a significant increase in LA3.0 a-wave latency ($p < 0.05$) in bipolar patients compared to controls. In contrast, the Flicker P2-wave amplitude is significantly higher ($p < 0.01$) in controls. We extracted 12 significant DWT coefficients, 7 in ERGs and 5 in EEGs while we had only 3 features in the time domain.

We obtained better classification results using TF features rather than temporal characteristics for any electrode, whether they are coupled or not, as shown in Table I.

TABLE I. SCORES (MEAN (SD)) FOR COUPLED AND NON COUPLED ERG AND EEG FEATURES

Electrode	Feature	$F1_{score}$	Accuracy	Recall	Specificity
EEG	Amp./Lat.	65.4 (12.8)	60.2 (11.3)	72.7 (21)	45.2 (20.5)
	DWT	75.5 (12.3)	73.1 (14.0)	76.7 (15.8)	68.8 (21.8)
ERG	Amp./Lat.	70.9 (10.1)	67.5 (11.3)	73.3 (13.9)	60.4 (20.4)
	DWT	76.5 (11.4)	74.4 (10.3)	79.7 (17.3)	68.0 (15.1)
EEG/ERG	Amp./Lat.	74.4 (9.6)	68.4 (11.5)	84.7 (14.2)	48.8 (20.7)
	DWT	82.8 (9.2)	80.4 (10.1)	87.3 (12.9)	72.0 (15.7)

Moreover, we show that combining EEG and ERG yields in greater scores with a decrease in the variability for most of the scores despite high standard deviations for EEG.

Finally, coupled EEG-ERG TF showed the best results with a high recall ($> 87\%$) meaning that a few bipolar patients will remain undiagnosed, whereas the specificity is lower (72%).

IV. Conclusion and Future Work

Our first results suggest that the TF features give a more precise representation of the ERG and EEG signals compared to the amplitudes and latencies of the waves. They also suggest that coupled ERG/EEG provides greater discrimination and more reliable predictions, making it highly beneficial for BD diagnosis. However, the relatively small data set might limit the generalizability of the obtained results. Our future work will focus on improving these results by including more flash stimuli and testing other machine learning classifiers.

References

[1] M. Berk *et al.*, "Setting the stage: From prodrome to treatment resistance in bipolar disorder", *Bipolar Disorders*, vol. 9, no. 7, pp. 671–678, 2007.

[2] S. M. Silverstein, D. L. Demmin, J. B. Schallek, and S. I. Fradkin, "Measures of retinal structure and function as biomarkers in neurology and psychiatry", *Biomarkers in Neuropsychiatry*, vol. 2, p. 100 018, 2020.

[3] A. Tan, T. Schwitzer, J. Conart, and K. Angioi-Duprez, "Study of retinal structure and function in patients with major depressive disorder, bipolar disorder or schizophrenia: A review of the literature", *Journal Français d'Ophtalmologie*, vol. 43, no. 5, e157–e166, 2020.

[4] T. Schwitzer, J. Lavoie, A. Giersch, R. Schwan, and V. Laprevote, "The emerging field of retinal electrophysiological measurements in psychiatric research: A review of the findings and the perspectives in major depressive disorder", *Journal of Psychiatric Research*, vol. 70, pp. 113–120, 2015.

[5] M. Hébert *et al.*, "The electroretinogram may differentiate schizophrenia from bipolar disorder", *Biological Psychiatry*, vol. 87, no. 3, pp. 263–270, 2020.

[6] T. Schwitzer *et al.*, "Retinal electroretinogram features can detect depression state and treatment response in adults: A machine learning approach", *Journal of Affective Disorders*, vol. 306, pp. 208–214, 2022.

[7] E. Bubl *et al.*, "Retinal dysfunction of contrast processing in major depression also apparent in cortical activity", *European Archives of Psychiatry and Clinical Neuroscience*, vol. 265, no. 4, pp. 343–350, Jun. 2015.

[8] K. Tursini *et al.*, "Subsequent and simultaneous electrophysiological investigation of the retina and the visual cortex in neurodegenerative and psychiatric diseases: What are the forecasts for the medicine of tomorrow?", *Frontiers in Psychiatry*, vol. 14, p. 1 167 654, 2023.

[9] P. B. Patil and M. S. Chavan, "A wavelet based method for denoising of biomedical signal", in *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)*, 2012, pp. 278–283.

[10] J. V. Odom *et al.*, "ISCEV standard for clinical visual evoked potentials: (2016 update)", *Documenta Ophthalmologica*, vol. 133, no. 1, pp. 1–9, 2016.

[11] A. G. Robson *et al.*, "ISCEV standard for full-field clinical electroretinography (2022 update)", *Documenta Ophthalmologica*, vol. 144, no. 3, pp. 165–177, 2022.

[12] D. A. Pisner and D. M. Schnyer, "Chapter 6 - support vector machine", in *Machine Learning*, A. Mechelli and S. Vieira, Eds., Academic Press, 2020, pp. 101–121.

A New 1D-CNN Paradigm for Onset Detection of Absence Seizures in Children

Maxime Yochum[1], Amar Kachenoura[1], Matthieu Aud'hui[1], Fabrice Wendling[1], Anna Kaminska[2], Rima Nabbout[2], Mathieu Kuchenbuch[3], Pascal Benquet[1]

[1] University of Rennes
INSERM, LTSI-U1099
35000 Rennes, France
maxime.yochum@univ-rennes.fr
amar.kachenoura@univ-rennes.fr
matthieu.audhui@univ-rennes.fr
fabrice.wendling@inserm.fr
pascal.benquet@univ-rennes.fr

[2] Hopital Necker Enfants Malades,
Institute Imagine INSERM 1163,
Universite de Paris, Paris, France.
anna.kamins@aphp.fr
rima.nabbout@aphp.fr

[3] Pediatric and Genetic Department
CHU , Nancy, France
m.kuchenbuch@chru-nancy.fr

***Abstract*—This study presents a One-Dimensional Convolutional Neural Network (1D-CNN)-based algorithm for the early detection of childhood absence seizures in ElectroEncephaloGraphy (EEG) traces. This detection aims to enable timely sensory interventions, such as acoustic or visual stimulation, to potentially abort seizures. The algorithm was evaluated using a clinical setting with full EEG data and a reduced number of electrodes version of the data to show its suitability in a normal child environment. On the clinical EEG database of 117 patients, the model achieved promising results, including a Sensitivity of 0.859, Precision of 0.819, F1-score of 0.837, and a mean detection delay of 0.522 seconds. The performance remained satisfactory when using fewer electrodes, with a Sensitivity of 0.837, Precision of 0.808, F1-score of 0.820, and similar detection delays. These results demonstrate the method's robustness and feasibility for clinical applications, as well as its potential to be embedded in wearable devices for continuous, real-time seizure monitoring and intervention in children with absence epilepsy.**

Keywords-Surface EEG; Childhood Absence Epilepsy; Onset Detection; 1D-CNN.

I. INTRODUCTION

Typical absence seizures are characterized by brief and sudden lapses in consciousness and an absence of voluntary movements. Typically, they are associated with specific patterns of generalized spike-wave discharges observed in EEG recordings [1]. Childhood Absence Epilepsy (CAE) affects between 6.3 to 8.0 children per 100 000 annually [2] and accounts for 18% of all cases of epilepsy in school-aged children. Absence seizures, if untreated, can occur frequently throughout the day, sometimes up to 200 episodes daily [3]. Children with CAE generally follow a normal developmental path. Nevertheless, approximately 30% of them experience learning difficulties and Attention Deficit Hyperactivity Disorder (ADHD).

The diagnosis of CAE often involves a physical exam with an EEG routine during voluntary hyperventilation. On EEG, seizures are commonly associated with 3-4 Hz generalized spike-wave patterns, but variations in speed, symmetry, and the presence of polyspikes may also be observed. The treatment of absence seizures typically involves antiepileptic drugs, although there is a notable drug resistance rate of approximately 30% [4]. As an alternative to drug therapy, sensory or electrical stimulation techniques have shown promise in interrupting seizures [5], [6]. Research on rodent models has demonstrated that auditory stimuli, such as a 2 kHz tone, during the first few seconds of the seizure can stop around 52% of absence seizures [6]. In humans[7], simple acoustic stimuli delivered during the first 3 seconds of the seizure can inhibit the episode with a success rate of 57%. Thus, detecting the onset of seizures as early as possible is crucial for effectively applying these kinds of stimulation techniques. Numerous studies for absence seizure detection from surface EEG signals have been reported during the last decades [8]–[12]. Surprisingly, none of these approaches have focused on early detection, i.e., identifying seizure within one second of its onset, which is crucial for applying external stimulation to abort seizures as the stimulation must occur within the first seconds of the seizure [7]. To address this gap, the paper proposes a new Deep Learning-based (DL) approach designed for early detection of absence seizures from raw EEG data. The method uses a learned 1D-CNN model to identify seizure onset within short sliding windows of EEG data in real-time, making it suitable for integration into wearable devices. This approach improves accuracy by analyzing data across multiple EEG channels. Additional constraints on the consecutive detection of the onset of seizures and the number of channels, where the seizure is detected, are also proposed to minimize the False Detection Rate (FDR), ensuring robustness of the pipeline in real-world applications.

This communication is organized as follows: the dataset, the CNN-based model and the evaluation criteria are presented in Section II. The obtained results are reported in Section III. Discussion, conclusions and perspectives are given in Section IV.

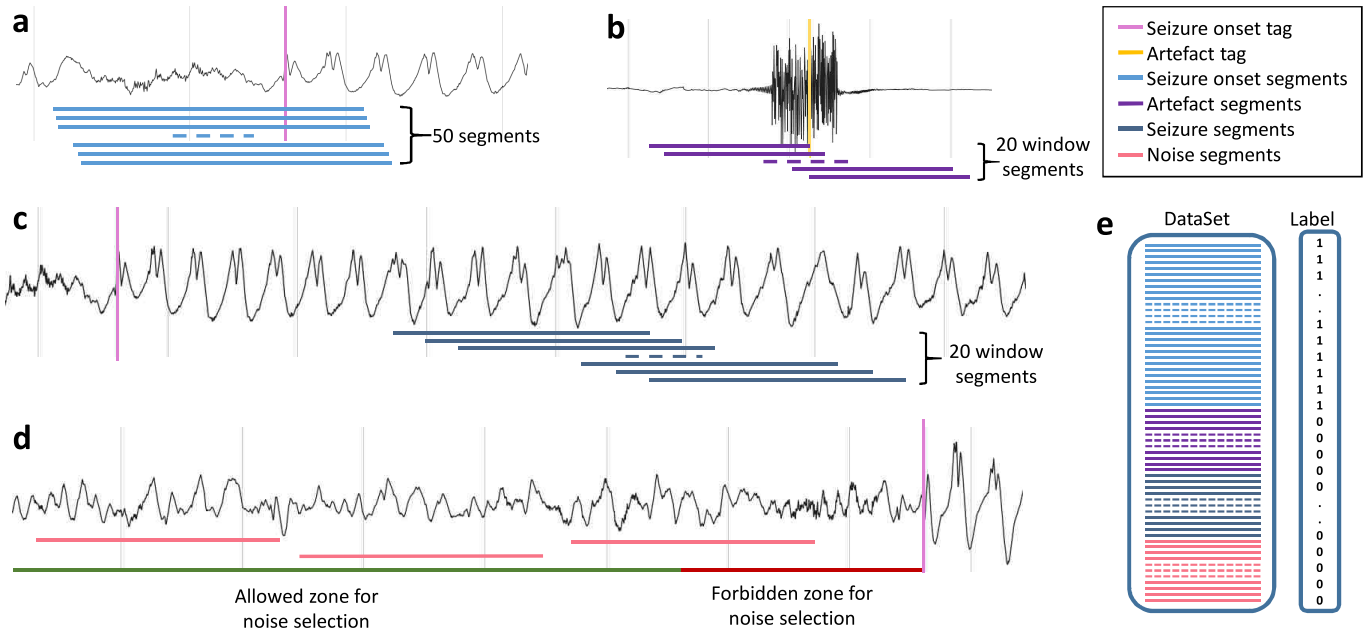


Figure 1. a. Selection of 50 segments for the seizure onset. The first segment is selected so the expert onset tag is located at the 384^{th} sample. The other segments are shifted from the first one from 1 to 49 samples. b. 20 segments were picked from -2 to 2s around the artifact tag (yellow position). c. 20 seizure segments were picked starting from 2s to 4s after the seizure onset tag. d. Noise segments were picked where seizure onset tags were absent within 2s from the starting noise segment.

II. Material and Method

The annotated dataset, the data-driven model design and the evaluation metrics are described in this section.

A. EEG recordings

In this study, EEG signals issued from 117 children (53 females and 64 males) diagnosed with CAE were used for evaluating the proposed pipeline. The dataset was acquired between 2013 and 2019, following the guidelines outlined in the French recommendations for EEG procedures in children [13] under the study protocol IRB:IORG0010044. The children were between 4 and 11 years old, and the recordings were conducted at two medical centers: Saint-Brieuc Hospital and Necker-Enfants Malades Hospital. The study strictly excluded children with intellectual disabilities or relevant neurological abnormalities based on the new classification of epileptic syndromes. EEG signals were acquired using the Deltamed Natus system at 256 Hz sampling frequency, with recordings lasting at least 20 minutes. The number of EEG electrodes varied across recordings, depending on the age of the patients, with 11, 16, or 19 electrodes used. Following the 10/20 international system, these recordings resulted in a total duration of 2.75 days of EEG data, or 49.9 days when measured across one EEG channel. As the signals are z-score normalized for each EEG trace, no magnitude scale was given in all figures.

B. EEGs annotation

It is well-known that the ground truth is mandatory for the performance evaluation of machine learning methods. In our study, clinical experts visually annotated the seizure onset times to create a ground truth for training the model and validate the detection of the proposed procedure. The experts used dedicated software to mark seizure onset times across each EEG channel in a recording. To ensure consistency, two strict criteria were applied for selecting seizure events: (1) at least four consecutive spike-wave occurrences had to be visually detected, and (2) spike-waves had to be visible on at least half of the EEG channels. This ensured that only generalized seizures were included in the analysis, leading to 827 early seizure onset positions used for training and testing.

C. Training data set building strategy

An adequate design of the training data set is important to construct an efficient and stable DL model. Thus, to address the specific task of early detection of absence seizure onset, the training set was built by dividing the EEG data into two sets of 2-second segments: the first one contains seizure onset segments and the second one encompasses non-seizure onset segments. More precisely, as depicted in Figure 1-a, the seizure onset set was constructed, by extracting 50 segments from each onset expert tags. These segments were designed to capture temporal information around the seizure onset by varying their relative position to the expert onset. This allows the model to learn the dynamic transition from background EEG to seizure activity. The seizure onset expert tags, positioned around 1.5 seconds of the window, ensure the presence of 1.5 to 2 spike waves at the end of the seizure onset segments which contributed to a comprehensive analysis of onset seizure events. Regarding the non-seizure onset set, it includes three subcategories of EEG signals: background

EEG, physiological and non-physiological artifacts, and fully developed seizure segments (Figure 1-b). Background EEG was randomly selected to represent a broad spectrum of normal brain activity (Figure 1-d). Artifact segments, such as those caused by patient movements, eye movements, or amplifier disconnections, were included to avoid detecting them as false positives (Figure 1-b). In addition, fully developed seizure segments (Figure 1-c) were also added to ensure that the model could differentiate between the onset of a seizure and the more periodic, established spike-wave patterns of a full seizure.

D. DL-based model architecture

The proposed model is designed to analyze each EEG electrode independently. Analyzing each EEG channel independently ensures flexibility across different EEG systems and configurations, making it suitable for various clinical settings. The model consists of four 1D convolutional layers with progressively increasing numbers of filters (from 32 to 256), followed by average pooling layers, a flatten layer, and two fully connected layers. A dropout rate of 50% was applied to prevent overfitting, and the Rectified Linear Unit (ReLU) activation function was used throughout the network. The final output layer used a Softmax activation function to classify segments as either seizure onset or non-seizure onset. The training was optimized using the Adam algorithm, with a batch size of 128, 10 epochs, and a learning rate of 0.001.

E. Training, detection stages

To ensure the generalizability, robustness, and stability of the proposed DL-based method, the training stage involved constructing 12 bootstrap datasets, with 80% of patients allocated for training and 20% for testing. Importantly, the model was trained based on a non-patient-specific detection strategy. For each bootstrap, patients included in the training set were excluded from the testing set. During the detection phase (testing stage), for each tested patient, the trained model was applied on each EEG channel using a 2-second sliding window with a 1-sample shift. Segments were classified as Event of Interest (EoI) based on the output probability of the 1D-CNN exceeding a threshold T. However, the sliding window approach could lead to multiple detections of the same seizure onset, artificially exaggerating the FDR. To reduce this issue, a post-processing step was introduced. It is based on two thresholds: i) if the percentage of the number of positive detections within the N consecutive 2 s time windows is higher than a threshold Pw%, then the final sample of the last window is qualified to be a seizure onset position, and ii) the end of this last window is definitively tagged as a seizure onset if it was simultaneously detected on a minimum percentage of EEG channels (denoted as Pch%).

F. Evaluation metrics

In this study, the Sensitivity (S), Precision (P), F1-score and FDR per Hour (FDR/H) metrics [14] are used to evaluate the seizure onset detection performance of the proposed pipeline. The limit for the detection was fixed to 2 seconds from an expert tag: if a detection of our algorithm is out of this ±2s bound, it is considered as a False Positive (FP).

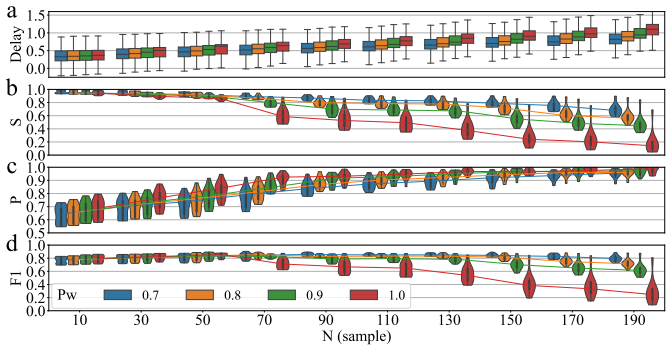


Figure 2. All plots x-axis represent the length (in sample) of the consecutive window use along with the Pw threshold, In all plots, each color used represents a Pw value (blue: Pw=70%; orange: Pw=80%; green: Pw=90%; red: Pw=100%). a. Boxplot of the delays of the algorithm detection with respect to the expert tags. b-d. violin plot of the Sensitivity (b), Precision (c) and F1-score (d).

III. RESULTS

The first experiment was conducted to determine a good compromise between the number N of the consecutive 2 s time windows and the threshold Pw. Regarding the Pch (minimum number of channels where the onset was simultaneously detected), it was fixed to 50% in the sequel. Figures 2 (a), (b), (c) and (d) display the delays, in seconds, of the algorithm detections relative to the expert annotations, S, P and F1-score, respectively, for all tested patients (across the 12 bootstraps). Four values of Pw=70% (blue), 80%(orange), 90% (green),100% (red) were tested, where the number N was varied from 10 to 190 with a step of 20. It can be seen from Figure 2 (a) that as N increases, the detection delay becomes more pronounced, regardless Pw. Figure 2 (b) shows that Sensitivity decreases significantly for $N > 50$, while Precision increases with increasing N, indicating a reduction in false detection. Interestingly, the best F1-score, defined as the harmonic average of the Sensitivity and the Precision, was obtained for $N = 50$ and Pw=80%, with a satisfactory detection delay around 0.5 s. Figure 3 focuses on the results obtained for Pw=80%, $N = 50$, and Pch=50% across all bootstraps. The average F1-score across all bootstraps was 0.837 ± 0.032, reflecting the model's effectiveness in detecting seizure onsets. Sensitivity and Precision were also well-balanced, with averages of 0.859 ± 0.030 and 0.819 ± 0.064, respectively, while the FDR/H remained low at 1.78 ±0.49. Furthermore, the delay between the detected seizure onsets and expert annotations was minimal, with an average delay of 0.522 seconds and a maximum delay of 1.5 seconds, as shown in Figure 3 (b). These results confirm the model's ability for very early seizure detection. Additionally, the proposed pipeline demonstrated robustness across different training and testing sets, since the standard deviations were low, whatever the analyzed metric.

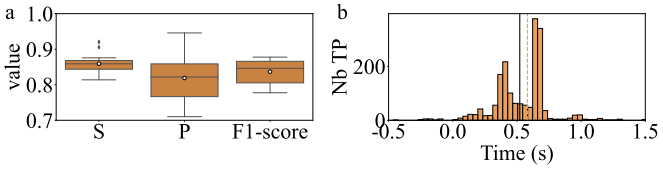


Figure 3. a: Sensitivity (S), Precision (P) and F1-score computed from the best F1 score of each Bootstrap (in blue) and for a unique triplet (in orange), means are shown with white circles. Histogram of delays measured between the expert annotation and the detected seizure onset moments by our algorithm for separate bootstrap optimization (b) and overall bootstrap triplet (c).

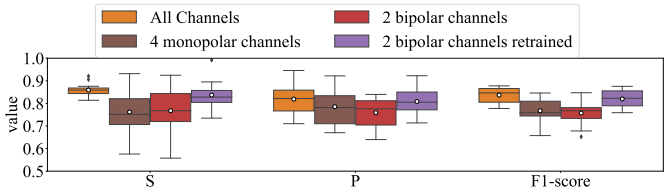


Figure 4. Sensitivity (S), Precision (P) and F1-score computed from the best F1 score for an overall triplet in four different cases: All channels (in orange), 4 monopolar channels (in green), 2 bipolar channels (in red), 2 bipolar channels with a retrained DL model (in purple). Medians are shown with black lines and means are shown with white circles

The second experiment deals with the configuration of the wearable device, where we can only expect that four electrodes will be available. Thus, we evaluated our detector only with Fp1, Fp2, T3 and T4 electrodes. The choice of two prefrontal electrodes and two temporal electrodes was driven by the fact that they could be hidden in the temples of glasses. More precisely, the model used previously was applied in two different montages: i) on 4 monopolar EEG channels (brown in Figure 4), and ii) on two bipolar channels Fp1-T3 and Fp2-T4 (red in Figure 4). Bipolar montages are known to be less susceptible to artifacts and commonly used for clinical EEG recordings. The impaired statistics using a reduced number of EEG channels are presented in Figure 4. We observed that the optimal F1-scores and the related sensitivities and precisions decrease for both montages compared to the use of all electrodes. Logically, the number of FDR/H increased from 1.783 for all electrodes to 3.071 and 3.060 for four monopolar and two bipolar electrodes, respectively.

To enhance the applicability of the model to bipolar channels, we also evaluate a new model that was specifically trained only on bipolar channels FP1-T3 and FP2-T4. As expected, this adjustment led to a significant improvement in results, although it does not exactly reach the performance achieved using all EEG channels (purple vs orange boxplots in Figure 4). With respect to the Performances of the initial model applied on bipolar montage (red boxplots), the Sensitivity, Precision and F1-score were increased from 0.78 to 0. 837 (±0. 064), 0.771 to 0. 808(±0. 063), and 0.796 to 0. 820 (±0. 040), respectively. In addition, the FDR/H was improved from 3.06 to 2.03. Regarding the delays of the detection of the seizure onset, the mean delay was almost not impacted (0.460 s).

IV. Discussion an conclusion

The proposed study is based on existing research, which demonstrates that absence seizures can be inhibited if external sensory stimulation is applied early in the seizure onset. Detection of the onset of the absence seizure as early as possible is mandatory to abort seizure progression, as delayed stimulation becomes ineffective once the seizure is fully established. Although several studies have been dedicated to the automated detection of absence seizures, no technique has yet been designed for early seizure onset detection (less than one second from the onset). Thus, this study introduces a new 1D-CNN-based pipeline for the early detection of absence seizures in children. Furthermore, the pipeline did not need heavy preprocessing and can be implemented in wearable devices. The 1D-CNN was favored over other models, such as Long Short Term Memory (LSTM) and Temporal Convolutional Network (TCN), due to its simplicity, ease of parallelization, and performance efficiency in handling EEG data. For instance, the computational time for processing data from a 15-electrodes is only about 0.4 ms. Obtained results, on a large real database, show that the model is very efficient in detecting the onset of seizures in children, with a Sensitivity of 0.859, Precision of 0.819, and F1-score of 0.837, alongside a time delay of just 0.522 seconds from the expert annotations. Importantly, even with a reduced set of electrodes (two bipolar channels), the method maintained good performance, which indicates that the algorithm is well-suited for portable devices. An adjustment of some parameters in the postprocessing step can also provide a possibility for a tradeoff between FDR/H and the maximal delay of detection allowed by a physicist to abort seizures.

The study acknowledges certain limitations, including the challenge of dealing with false detection due to short spike trains, which clinicians do not consider as seizures. In addition, more intensive clinical or animal studies are necessary to determine the optimal window length for effective intervention. The exploited EEG data were collected in controlled environments, and future work should focus on validating the robustness of the proposed pipeline in more variable settings, particularly in wearable devices.

Acknowledgment

This study has been funded by the Institut des Neurosciences Cliniques de Rennes (INCR, www.incr.fr), as part of the PREDILEPSY Project, and the Agence nationale de la recherche (ANR) Recherche Hospitalo-Universitaire en santé (RHU) through the innov4-epik project.

References

[1] V. Crunelli *et al.*, “Clinical and experimental insight into pathophysiology, comorbidity and therapy of absence seizures,” *Brain*, vol. 143, no. 8, pp. 2341–2368, 2020.

[2] E. Hirsch *et al.*, “Ilae definition of the idiopathic generalized epilepsy syndromes: Position statement by the ilae task force on nosology and definitions,” *Epilepsia*, vol. 63, no. 6, pp. 1475–1499, 2022.

[3] M. R. de Feo, O. Mecarelli, G. Ricci, and M. F. Rina, "The utility of ambulatory eeg monitoring in typical absence seizures," *Brain and Development*, vol. 13, no. 4, pp. 223–227, 1991.

[4] M. Koutroumanidis *et al.*, "The role of eeg in the diagnosis and classification of the epilepsy syndromes: A tool for clinical practice by the ilae neurophysiology task force (part 1)," *Epileptic Disorders*, vol. 19, no. 3, pp. 233–298, 2017.

[5] H. Blumenfeld, "Consciousness and epilepsy: Why are patients with absence seizures absent?" *Progress in brain research*, vol. 150, pp. 271–603, 2005.

[6] S. Saillet *et al.*, "Neural adaptation to responsive stimulation: A comparison of auditory and deep brain stimulation in a rat model of absence epilepsy," *Brain stimulation*, vol. 6, no. 3, pp. 241–247, 2013.

[7] P. Rajna and C. Lona, "Sensory stimulation for inhibition of epileptic seizures," *Epilepsia*, vol. 30, no. 2, pp. 168–174, 1989.

[8] J. Duun-Henriksen *et al.*, "Automatic detection of childhood absence epilepsy seizures: Toward a monitoring device," *Pediatric neurology*, vol. 46, no. 5, pp. 287–292, 2012.

[9] P. Glaba *et al.*, "Absence seizure detection algorithm for portable eeg devices," *Frontiers in Neurology*, vol. 12, p. 685 814, 2021.

[10] T. W. Kjaer, H. B. Sorensen, S. Groenborg, C. R. Pedersen, and J. Duun-Henriksen, "Detection of paroxysms in long-term, single-channel eeg-monitoring of patients with typical absence seizures," *IEEE journal of translational engineering in health and medicine*, vol. 5, pp. 1–8, 2017.

[11] L. Swinnen *et al.*, "Accurate detection of typical absence seizures in adults and children using a two-channel electroencephalographic wearable behind the ears," *Epilepsia*, vol. 62, no. 11, pp. 2741–2752, 2021.

[12] C. Chatzichristos *et al.*, "Multimodal detection of typical absence seizures in home environment with wearable electrodes," *Frontiers in Signal Processing*, vol. 2, p. 1 014 700, 2022.

[13] A. Kaminska, F. Cheliout-Heraut, M. Eisermann, A. T. de Villepin, and M. Lamblin, "Eeg in children, in the laboratory or at the patient's bedside," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 45, no. 1, pp. 65–74, 2015.

[14] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation," in *Australasian joint conference on artificial intelligence*, Springer, 2006, pp. 1015–1021.

An Integrative Strategy for Solving the EEG Inverse Problem and the Estimation of Brain Effective Connectivity in Epilepsy. A Proof-of-Concept Study.

Marc Greige Ahmad Karfoul Pascal Benquet Maxime Yochum Régine Le Bouquin Jeannès
Univ Rennes, Inserm, LTSI - UMR 1099, F-35000 Rennes, France
email: marc.greige@univ-rennes.fr ahmad.karfoul@univ-rennes.fr pascal.benquet@univ-rennes.fr
maxime.yochum@univ-rennes.fr regine.le-bouquin-jeannes@univ-rennes.fr

***Abstract*—This study presents an integrative strategy for simultaneously localizing brain sources and inferring effective connectivity. The proposed approach leverages the model underlying the events of interest as a regularizer in the electroencephalographic inverse problem. The effectiveness of this strategy is confirmed using realistic simulated high resolution electroencephalographic signals in the context of epilepsy, and compared to the conventional sequential strategy, where connectivity estimation is performed after solving the electroencephalographic inverse problem.**

Keywords- Effective connectivity; inverse problem; optimization; EEG; epilepsy.

I. Introduction

Inferring effective connectivity among brain regions from surface High Resolution (HR)-ElectroEncephaloGraphic (EEG) recordings is typically performed through a sequential process involving: (i) preprocessing EEG data to remove artifacts and detect Events of Interest (EIs); (ii) solving the EEG inverse problem to pinpoint the spatial location of brain regions responsible for the observed EIs and reconstruct their neural activities; and (iii) inferring effective connectivity among these identified regions based on their reconstructed neural activities. However, this sequential approach faces two major limitations: (i) the error propagation phenomenon across the successive steps, and (ii) the absence of the optimal pairing of source localization methods and effective connectivity measures. Furthermore, even if such an optimal pairing exists, it is highly dependent on the specific application. To address these limitations, a Proof-of-Concept (PoC) study of an integrative strategy that combines both source localization and brain effective connectivity inference steps into a single one is presented here. Note that a similar strategy was recently proposed in [1] but within the context of functional connectivity. The proposed integrative strategy is evaluated here in the context of drug-resistant epilepsy [2], where identifying brain connectivity is essential for localizing regions responsible for seizure initiation and propagation. This information is valuable for surgical treatments aimed at reducing or eliminating seizures. The performance of the proposed integrative strategy is tested using realistic simulated HR-EEG signals and compared to the conventional sequential strategy, where brain connectivity is determined after solving the EEG inverse problem. The remainder of this paper is organized as follows: Section II details the proposed integrative strategy, including the EEG observation model and optimization framework. Section III presents simulation results, comparing its performance to the conventional sequential method. Finally, Section IV summarizes the findings, discusses implications for epilepsy research, and suggests future directions.

II. Towards an Integrative Strategy

This section presents the concept of the proposed integrative strategy, emphasizing the key idea of combining the brain source localization problem with brain effective connectivity inference.

A. EEG observation model

From now on, HR-EEG recordings are assumed to be preprocessed, with artifacts removed and EOIs (pre-ictal epileptic spikes) detected. Assume that the brain is divided into P regions, each consisting of synchronized dipoles in the source space. Then, the brain electrical activity over T time points, observed by N scalp EEG sensors, follows the linear model:

$$\boldsymbol{X} = \boldsymbol{G}\boldsymbol{Y} + \boldsymbol{X}_b$$

where $\boldsymbol{X} \in \mathbb{R}^{N\times T}$ is the spatio-temporal observation matrix, $\boldsymbol{G} \in \mathbb{R}^{N\times P}$ is the lead field matrix, which is a known matrix encoding the transfer medium between the cortical surface (source space) and the scalp (observation space), $\boldsymbol{Y} \in \mathbb{R}^{P\times T}$ collects the neural activities of epileptic regions, and $\boldsymbol{X}_b \in \mathbb{R}^{N\times T}$ corresponds to background brain activity.

B. EEG inverse problem

The EEG inverse problem involves estimating the positions of brain sources underlying the EIs ($e.g.$, pre-ictal epileptic spikes) and reconstructing their corresponding electrical activities. To this end, the following optimization problem is to be solved:

$$\underset{\boldsymbol{Y}}{\text{Minimize}}\ ||\boldsymbol{X} - \boldsymbol{G}\boldsymbol{Y}||_{\text{F}}^2 + \sum_{c=1}^{C} \lambda_c f_c(\boldsymbol{Y}) \qquad (1)$$

Here, f_c represents the c-th regularization term, encoding prior information about the latent source matrix $\boldsymbol{Y}$, $\lambda_c \in \mathbb{R}_+^*$ is the associated penalty parameter, and $\|.\|_{\text{F}}$ is the Frobenius norm. For example, in the Weighted Minimum Norm Estimate (wMNE) approach [3][4], widely used to solve the EEG inverse problem for its simplicity and efficiency, the regularization term is $f_1(\boldsymbol{Y}) = \|\boldsymbol{B}\boldsymbol{Y}\|_{\text{F}}^2$, where $\boldsymbol{B}$ is a weighting matrix with diagonal entries $B_{p,p} = \|\boldsymbol{g}_p\|_2^{-1}$. Here, $\boldsymbol{g}_p$ denotes the p-th column of $\boldsymbol{G} \in \mathbb{R}^{N\times P}$. The role of $\boldsymbol{B}$ is to compensate for the bias in the estimation of deep brain sources.

C. The proposed integrative strategy

As previously mentioned, the proposed integrative strategy unifies source localization and brain effective connectivity inference into a single step. In the context of epilepsy, pre-ictal epileptic spikes, events occurring just before seizure onset, offer valuable insights into the brain regions initiating seizures. The key idea of the integrative strategy is to incorporate the mathematical model underlying the EOIs as an additional regularization term in the EEG inverse problem. In this PoC study, a MultiVariate AutoRegressive (MVAR) model is employed to describe the pre-ictal epileptic spikes. Albeit suboptimal as neural activities exhibit rather nonlinear interactions, the MVAR model is widely adopted in effective connectivity measures ($e.g.$, Granger index [5][6]). An MVAR modeling of $\boldsymbol{Y}$ is given by:

$$\boldsymbol{Y} = \sum_{l=1}^{L} \boldsymbol{\Theta}^l \boldsymbol{Y}^l + \boldsymbol{W} \tag{2}$$

where $\boldsymbol{\Theta}^l \in \mathbb{R}^{P\times P}$ denotes the matrix of model coefficients, $\boldsymbol{Y}^l \in \mathbb{R}^{P\times T}$ is a delayed version of $\boldsymbol{Y} \in \mathbb{R}^{P\times T}$ associated with the l-th time lag, and $\boldsymbol{W}$ accounts for the model residual, where the (i,j)-th entry of $\boldsymbol{W}$ verifies $W_{i,j} \sim \mathcal{N}(0,\sigma)$. The elements of the L matrices $\boldsymbol{\Theta}^l$ reflect, to a large extent, causal effects that those delayed signals have on the signal they are constituting. Thus, estimating these coefficients will lead to infer the causal relationships among different epileptic sources. Now, by considering the well-known wMNE algorithm for source localization and the MVAR model as a model underlying the observed pre-ictal epileptic spikes, the proposed integrative strategy consists in solving the following optimization problem:

$$\underset{\boldsymbol{Y},\{\boldsymbol{Y}^l\}_{1\le l\le L},\{\boldsymbol{\Theta}^l\}_{1\le l\le L}}{\text{Minimize}} \; ||\boldsymbol{X}-\boldsymbol{G}\boldsymbol{Y}||_F^2+\gamma||\boldsymbol{Y}-\sum_{l=1}^{L}\boldsymbol{\Theta}^l\boldsymbol{Y}^l||_F^2 + \lambda||\boldsymbol{B}\boldsymbol{Y}||_F^2 + \xi\sum_{l=1}^{L}||\boldsymbol{B}\boldsymbol{Y}^l||_F^2 + \beta\sum_{l=1}^{L}||\boldsymbol{\Theta}^l||_1 \tag{3}$$

where γ, λ, ξ, and β are hyperparameters optimized using a grid search strategy to achieve the best results. The inclusion of the L_1-norm term emphasizes the selection of only the most significant connections among brain regions. Solving the above optimization problem is performed by minimizing instead its associated augmented Lagrangian function, where the Proximal Alternating Linearized Minimization (PALM) algorithm [7] is used as a solver.

III. Numerical Results

To assess the feasibility and performance of the proposed integrative strategy, a realistic simulated 257-channel HR-EEG dataset of 60 seconds with a sampling frequency of 1024 Hz was generated to model a focal epileptic seizure. In this simulation, the right frontal pole (r-FP) region was defined as the seizure onset zone, while the right middle temporal gyrus (r-MT) region represented the propagation zone, establishing a causal effect from r-FP to r-MT. The dataset was created using the "Coalia" software [8], which incorporates realistic head models. The brain was parcellated into 66 regions based on the Desikan-Killiany atlas [9]. As far as the regularization parameters γ, λ, ξ and β, were concerned, they were set to $0.5, 23, 1$ and 1, respectively. The proposed strategy was compared to the traditional sequential approach, where the wMNE algorithm was employed to solve the EEG inverse problem, followed by Granger causality [5] to estimate effective connectivity among the localized neural sources based on their reconstructed activities. For both strategies, the study was conducted over a time period of 6 seconds right before the onset of the epileptic seizure. In terms of source localization, the proposed integrative strategy demonstrated clear superiority over the sequential strategy based on the wMNE algorithm [3], [4], as illustrated in Figure 1.

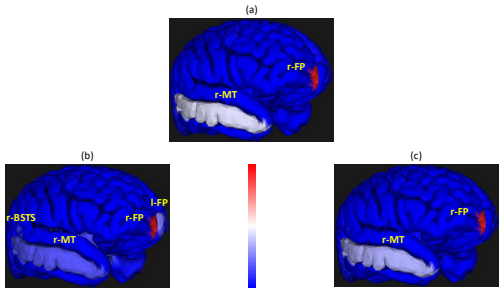


Figure 1. Epileptic source localization. (a) ground truth, (b) sequential strategy with wMNE, (c) proposed integrative strategy.

Specifically, in addition to correctly localizing the brain regions associated with the seizure (r-FP for the onset region and r-MT for the propagation region), the sequential strategy where the wMNE algorithm is used to solve the EEG inverse problem, followed by the Granger causality measure for inferring effective connectivity, also identified other spurious brain regions, such as the left frontal pole (l-FP) region as an onset region and the right banks of the superior temporal sulcus (r-BSTS) as a propagation region. In contrast, the proposed integrative strategy did not identify any spurious regions, providing a more accurate and reliable result. Both strategies successfully identified the correct causal effect between the two brain regions, r-FP and r-MT. This effect is highlighted in bold, as shown in Table I for the conventional sequential strategy and Table II for the proposed integrative strategy, where the interactions are ranked from highest (leftmost) to lowest (rightmost).

TABLE I
Estimated effective connectivity using the sequential strategy.

Interaction 1	**Interaction** 2	**Interaction** 3	**Interaction** 4
l-FP → r-MT	**r-FP → r-MT**	r-FP → r-BSTS	l-FP → r-BSTS

TABLE II
ESTIMATED EFFECTIVE CONNECTIVITY USING THE INTEGRATIVE STRATEGY.

Lag	Interaction 1	Interaction 2
3	**r-FP $\rightarrow$ r-MT**	$\times$
5	**r-FP $\rightarrow$ r-MT**	$\times$
6	$\times$	**r-FP $\rightarrow$ r-MT**
Average	**r-FP $\rightarrow$ r-MT**	

It is noteworthy that the connectivity matrices obtained for each strategy were thresholded such that all values in the matrices that were less than 90% of the largest value were set to zero. This thresholding step ensured that only the most significant connectivity relationships were retained for further analysis. However, compared to the sequential strategy, the proposed integrative strategy offers the possibility for a dynamic analysis of the effective connectivity over the different time lags. For some time lags ($i.e.$, $l \in \{3, 5\}$), the highest estimated causal interaction ($i.e.$, Interaction 1), corresponds to the true effective connectivity while for other time lags ($i.e.$, $l = 6$), it stands for the second most important connectivity value ($i.e.$, Interaction 2). Thus, contrary to the sequential strategy, where the true causal effect is ranked as the second most important connectivity pattern (see Table I), the integrative strategy offers an average effective connectivity over the considered time lags, where the true effective connectivity accounts for the most significant connectivity pattern (see Table II). It is noteworthy that Table II highlights only the interactions among the regions of interest (i.e., r-FP and r-MT).

IV. CONCLUSION AND FUTURE WORK

In this communication a PoC study of an integrative strategy for simultaneous brain source localization and effective connectivity estimation was proposed. The strategy relied mainly on the model underlying the EIs as an additional regularization term in the source localization problem. This strategy was evaluated in the context of focal epilepsy with pre-ictal epileptic spikes as EIs that were assumed to follow an MVAR model. The effectiveness of this integrative solution was confirmed using realistic surface HR-EEG recordings compared with the conventional sequential strategy, where wMNE was considered for source localization and Granger causality for effective connectivity estimation. Future work will focus on evaluating the proposed strategy in more complex scenarios, such as incorporating multiple epileptic sources and evaluating its performance on real HR-EEG data from multiple epileptic patients.

REFERENCES

[1] A. Karfoul, A. Kachenoura, and L. Albera, "A unified approach for inverse problem in EEG and brain connectivity with application to epilepsy. a proof of concept study", *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 1713–1717, 2023. DOI: 10.23919/EUSIPCO58844.2023.10289968.

[2] P. Kwan, S. C. Schachter, and M. J. Brodie, "Drug-resistant epilepsy", *New England Journal of Medicine*, vol. 365, no. 10, pp. 919–926, 2011.

[3] M. S. Hämäläinen, "Interpreting measured magnetic fields of the brain: Estimates of current distributions", *Helsinki University of Technology, Report*, 1984.

[4] R. D. Pascual-Marqui, C. M. Michel, and D. Lehmann, "Low resolution electromagnetic tomography: A new method for localizing electrical activity in the brain", *International Journal of Psychophysiology*, vol. 18, no. 1, pp. 49–65, 1994.

[5] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods", *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[6] B. He *et al.*, "Electrophysiological brain connectivity: Theory and implementation", *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 7, pp. 2115–2137, 2019. DOI: 10.1109/TBME.2019.2913928.

[7] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems", *Mathematical Programming*, vol. 146, pp. 459–494, 2014. DOI: 10.1007/s10107-013-0701-9.

[8] S. Bensaid, J. Modolo, I. Merlet, F. Wendling, and P. Benquet, "Coalia: A computational model of human eeg for consciousness research", *Frontiers in Systems Neuroscience*, vol. 13, p. 59, 2019.

[9] R. S. Desikan *et al.*, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest", *Neuroimage*, vol. 31, no. 3, pp. 968–980, 2006.

FMSTFnet: Feature-Modulation Spatio-Temporal Fusion Network for HDR Video Reconstruction

Wei Zhang
Faculty of Info. Sci. & Eng.
Ningbo University
Ningbo, China
350192287@qq.com

Yeyao Chen
Faculty of Info. Sci. & Eng.
Ningbo University
Ningbo, China
chenyeyao@nbu.edu.cn

Gangyi Jiang
Faculty of Info. Sci. & Eng.
Ningbo University
Ningbo, China
jiangggangyi@126.com

***Abstract*—This paper proposes a novel High Dynamic Range video (HDRv) reconstruction method from Standard Dynamic Range video (SDRv), with a Feature Modulation Spatio-Temporal Fusion network (FMSTFnet). FMSTFnet has low-frequency and high-frequency parts with a pyramid structure. The low-frequency part mainly includes a Combined Global and Local Feature Modulation module (CGLFM) and a Spatio-Temporal Fusion Module (STFM). CGLFM modulates global and local features of SDR frames to correct the detail deviation caused by brightness differences in different regions and obtain preliminary HDR frames. STFM is designed to enhance the preliminary HDR frames using inter-frame information, and eliminate possible inter-frame artifacts. Finally, an adaptive hybrid module is constructed to fuse the low-frequency HDR frames and gradually extend the processed high-frequency information from low resolution to the higher. The proposed network fully utilizes the inter-frame information of multiple SDR frames and the contextual information of previously predicted HDR frames to generate high-quality results that are consistent in the temporal domain. The experimental results show that compared with many representative methods, the proposed method can reconstruct higher quality HDR videos.**

Keywords-high dynamic range video reconstruction; feature modulation; spatio-temporal fusion; transformer block.

I. INTRODUCTION

New generation displays can display visual contents with High Dynamic Range (HDR) and wide color gamut, providing a higher visual experience quality. However, at present, most video resources are still stored as Standard Dynamic Range videos (SDRv), resulting in a shortage of HDR video (HDRv) resources. Thus, generating HDRv from SDRv (SDRv-to-HDRv) is a challenging task [1][2].

For learning-based SDRv-to-HDRv, Kim et al. [3] proposed a method with separating input SDR frame into base and detail layers for different processing, which has the advantage of being easier to restore fine details. Subsequently, they integrated video super-resolution with SDRv-to-HDRv task to enhance texture details [4]. Chen et al. [5] designed a deep learning network for a single SDRv-to-HDRv task, which includes global feature modulation, local enhancement, and over-exposure compensation, and achieved good results. Wang et al. [6] proposed an SDRv-to-HDRv method with three sub-networks corresponding to the three processes in HDR imaging pipeline, to generate HDR images with rich global information. Xu et al. [7] constructed a frequency-aware modulation network that enhances the contrast of SDR to HDR conversion in a frequency adaptive manner, for reducing structural distortion and artifacts in the low-frequency regions. Xue et al. [8] proposed an improved residual block for extracting and fusing multi-layer features for fine-grained HDR image reconstruction. Guo et al. [1] constructed an HDRTV4K dataset and an HDR to SDR degradation model, and proposed a brightness segmentation network consisting of a global mapping backbone and two Transformer branches on the brightness range. The above methods mainly perform SDRv-to-HDRv tasks spatially. Many SDRv-to-HDRv methods mainly utilize a single SDR frame to generate corresponding HDR frame, which may lead to temporal inconsistency of HDRvs and produce annoying artifacts. Cao et al. [9] presented a kernel prediction network based SDRv-to-HDRv method, which utilizes multi-frame interaction modules to capture spatial information of multi-frame data and uses correction between adjacent frames to maintain inter-frame consistency.

In this paper, a novel SDRv-to-HDRv method with the design of Feature Modulation Spatio-Temporal Fusion network (FMSTFnet) is proposed. Its main contributions are summarized as follows: (1) A Combined Global and Local Feature Modulation module (CGLFM) is designed to perform macroscopic global and detailed local modulation on the current frame to reduce the color deviation of HDR video frames; (2) A Spatio-Temporal Fusion Module (STFM) is constructed, which can process contextual information in spatio-temporal domain, enhancing spatial results while reducing temporal inconsistencies. (3) Low-frequency and high-frequency information of SDRv are processed separately using a pyramid structure and fused with each other to obtain high-resolution output. Experimental results demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows. Section 2 describes the proposed method in detail, Section 3 gives experimental results and analyses, and finally Section 4 concludes the paper.

II. THE PROPOSED METHOD WITH FMSFNET

A novel SDRv-to-HDRv method with the designed FMSTFNet is proposed, as shown in Figure 1. Aiming at the problem of color deviation, a CGLFM is designed by

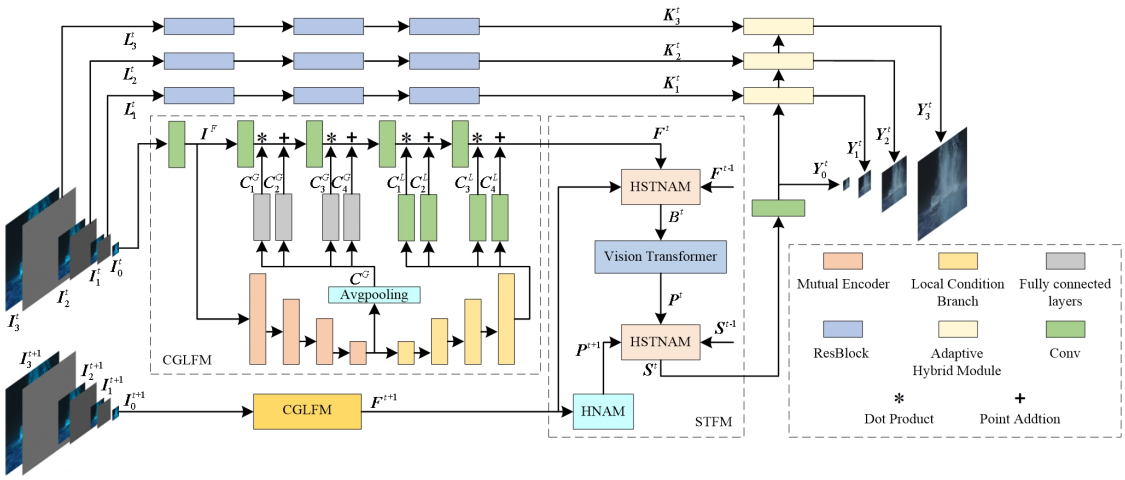


Figure 1. The proposed SDRv-to-HDRv method with the design of Feature Modulation Spatio-Temporal Fusion network (FMSTFnet).

combining adaptive feature modulation with Fourier convolution. For processing spatio-temporal information, a STFM is designed to fuse inter-frame features, and Transformer is employed to enhance the features, which can further reduce color deviation while eliminating temporal artifacts. The designed FMSFNet first establishes a pyramid structure and decomposes the input SDR frame into high-frequency component pyramids and low-frequency SDR frames. The low-frequency SDR frames are input to CGLFM and STFM to obtain low-frequency HDR frames. Residual blocks [10] are leveraged to reinforce the high-frequency components. Then, the enhanced high-frequency components are fused with low-frequency HDR frames using an Adaptive Hybrid Module (AHM), gradually expanding from low resolution to higher resolution results, and reconstructing the final high-resolution HDRv frame.

Specifically, for the t-th SDR frame $\boldsymbol{I}^t$, it is firstly decomposed into a Gaussian pyramid $\boldsymbol{M}_I^t = [\boldsymbol{I}_0^t, \boldsymbol{I}_1^t, \ldots \boldsymbol{I}_s^t]$ and a high-frequency component pyramid $\boldsymbol{M}_L^t = [\boldsymbol{L}_1^t, \ldots \boldsymbol{L}_s^t]$, where s is the number of downsampling. Similarly, $\boldsymbol{I}^{t+1}$ is also processed like $\boldsymbol{I}^t$. After that, the low-frequency components of $\boldsymbol{I}_0^t$ and $\boldsymbol{I}_0^{t+1}$ are respectively fed into CGLFM with weight sharing to obtain the preliminary HDR frames, denoted as $\boldsymbol{F}^t, \boldsymbol{F}^{t+1} = f_{CGLFM}(\boldsymbol{I}_0^t, \boldsymbol{I}_0^{t+1})$.

In Figure 1, $\boldsymbol{F}^t$ and $\boldsymbol{F}^{t+1}$ are then fed to STFM for spatio-temporal information enhancement; meanwhile, the (t-1)-th preliminary HDR frame $\boldsymbol{F}^{t-1}$ is also input to STFM to obtain the enhanced HDR frame $\boldsymbol{S}^t = f_{STFM}(\boldsymbol{F}^{t-1}, \boldsymbol{F}^t, \boldsymbol{F}^{t+1})$.

Each layer of the high-frequency component pyramid $\boldsymbol{M}_L^t$ is fed to multiple residual blocks $f_{Res}(\cdot)$, to enhance the high-frequency information, denoted as $\boldsymbol{K}_L^t = f_{Res}(\boldsymbol{M}_L^t)$. By relying on the high-frequency information $\boldsymbol{K}_L^t$ and the enhanced pyramid low-frequency HDR frame $\boldsymbol{S}^t$, high-resolution results can be reconstructed. Adaptive Hybrid Module (AHM) is used to fuse high-frequency component pyramids with low-frequency HDR frames, the final output pyramid $\boldsymbol{E}_L^t = [\boldsymbol{Y}_0^t, \boldsymbol{Y}_1^t, \ldots \boldsymbol{Y}_s^t]$ is obtained, where $\boldsymbol{Y}_s^t$ denotes the reconstructed HDR frames, $\boldsymbol{Y}_s^t = f_{AHM}(\boldsymbol{K}_L^t, \boldsymbol{S}^t)$.

A. CGLFM

In the SDRv-to-HDRv task, there may be a phenomenon of uneven repair of over-exposed and under-exposed regions, as well as uneven color mapping from standard color gamut to wide color gamut. To address this issues, CGLFM, as shown in Figure 1, is designed, in which the global rough modulation is for roughness adjustment on images, while the local detail fine-tuning is for local detail enhancement. Specifically, the input SDR frame $\boldsymbol{I}^t$ is processed through two-layer convolution to obtain low dynamic range features $\boldsymbol{I}^F$, which will be modulated into high dynamic range features $\boldsymbol{F}^t$. CGLFM has two parts, namely, conditional generation module and feature modulation module. The conditional generation module can extract global and local information from features for modulation. Global conditional generation module uses Fourier convolution to perform global operations on input features, and then uses average pooling to downsample while reducing information loss, so as to obtain global information of the image. After five downsampling and global pooling, the feature $\boldsymbol{C}^G$ is get, denoted by $\boldsymbol{C}^G = f_{AVG}(f_{CGFM}(\boldsymbol{I}^F))$, $f_{CGFM}(\cdot)$ and $f_{AVG}(\cdot)$ are the global operation and global pooling, respectively.

By processing $\boldsymbol{C}^G$, global conditional features $\boldsymbol{C}_V^G$ (V=A,B) are obtained, which are used as the global modulation vectors. Local modulation requires local features that represent the corresponding pixel positions in the image. Here, through upsampling the global features five times and decoding from the encoded global information, the local conditional features $\boldsymbol{C}_V^L$ is obtained and expressed by $\boldsymbol{C}_V^L = f_{CLFM}(f_{CGFM}(\boldsymbol{I}^F))$, and $f_{CLFM}(\cdot)$ is the local operation.

Then, perform global rough modulation and local detail fine-tuning on the features. The former uses global features $\boldsymbol{C}_A^G$ to point-multiply the SDR feature $\boldsymbol{H}_G$ to achieve global

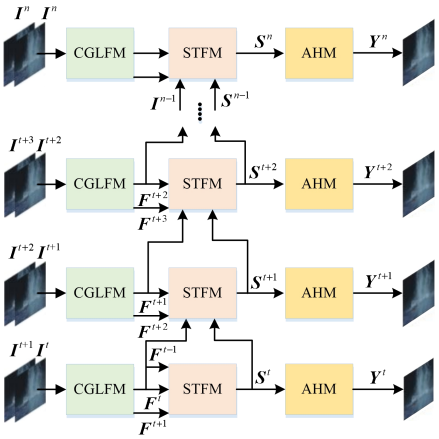
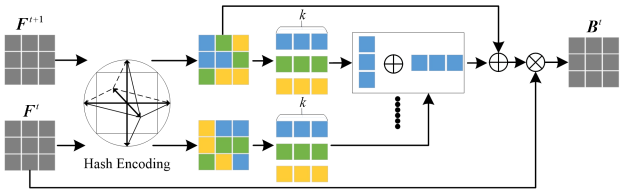
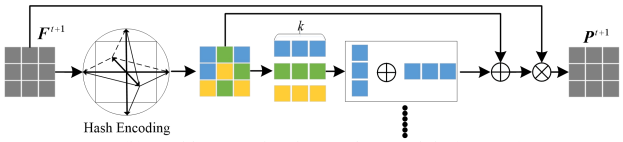


Figure 2. Information transmission approach of FMSTFNet.



(a) Hashing Spatio-Temporal Non-local Attention Module (HSTNAM)



(b) Hashing Non-local Attention Module (HNAM)

Figure 3. The used two non-local attention modules.

scaling, and directly adds $\boldsymbol{C}_B^G$ to achieve global displacement. The latter uses $\boldsymbol{C}_A^L$ to point-multiply the feature $\boldsymbol{H}_L$ to achieve local scaling, followed by adding $\boldsymbol{C}_B^L$ to achieve local displacement. After implementing local and global modulation, the features are converted to the HDR domain to obtain the preliminary HDR frame, which is expressed as

$$\boldsymbol{H}_L = \boldsymbol{C}_A^G * (\boldsymbol{H}_G) + \boldsymbol{C}_B^G \quad (1)$$

$$\boldsymbol{F}^t = \boldsymbol{C}_A^L * (\boldsymbol{H}_L) + \boldsymbol{C}_B^L \quad (2)$$

B. *Spatio-Temporal Fusion Module (STFM)*

STFM includes spatial and temporal reinforcement, mainly relying on the non-local attention mechanism. As shown in Figure 1, STFM mainly includes Hashing Spatio-Temporal Non-local Attention Module (HSTNAM), Hashing Non-local Attention Module (HNAM) [11], and Vision Transformer (ViT). To reduce resource consumption, when fusing inter-frame information in the temporal domain, only the information transmitted from the previous frame is used. Only the t-th and (t+1)-th frames are processed, and the (t-1)-th frame is obtained from the previous processing, as shown in Figure 2. Note that the (t-1)-th frame transmitted in the network is the intermediate feature rather than image. This processing can reduce the used memory while allowing the network to learn the entire sequence information. The input (t-1)-th frame contains the content of the previous video frames. As the number of input video frames increases, the network can learn all the early video frames.

STFM has four input features, i.e., $\boldsymbol{F}^t$, $\boldsymbol{F}^{t+1}$, $\boldsymbol{F}^{t-1}$ and $\boldsymbol{S}^{t-1}$. It has conducted two inter-frame information fusions, and with the deepening of the network, more deep level information is carried in the features. HSTNAM in Figure 3(a) is constructed to fuse the features of the t-th, (t-1)-th and (t+1)-th frames to obtain inter-frame information. Figure 3(b) represents the hashing non-local attention module, which differs from HSTNAM in that it only calculates spatial domain non-local attention. The purpose of STFM is to enhance features from both spatial and temporal perspectives, learn global inter-frame information to improve the temporal correlation of videos.

C. *HDR Reconstruction and Loss Function*

The FMSTFNet employs a pyramid structure, and the proposed method mainly focuses on handling the low-frequency components of the pyramid, which are processed using the above modules. For the high-frequency components, the stacked residual blocks are directly used for processing. AHM is constructed to facilitate rapid scaling of low resolution results. A lightweight module is designed as

$$\boldsymbol{Y}_{s+1}^t = h(\phi_2(cat(up(\phi_1(\boldsymbol{Y}_s^t)), \boldsymbol{K}_s^t))) \quad (3)$$

where $up(\cdot)$ is the bilinear interpolation, $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are two 3×3 convolutional layers, $cat(\cdot)$ is the channel concatenation, $\boldsymbol{K}_s^t$ is the high-frequency component of $\boldsymbol{I}^t$, and $h(\cdot)$ is the ReLU activation function.

The proposed loss function includes a multi-scale HDR reconstruction loss L_r and a perceptual loss L_p, expressed as

$$loss: L = \lambda_1 L_r + \lambda_2 L_p \quad (4)$$

where L_r represents the L_1 loss between the ground truth HDR image pyramid $\boldsymbol{H}_L$ and the predicted HDR image pyramid $\boldsymbol{Y}_L$. L_p is the L_1-norm difference between the intermediate feature maps when $\boldsymbol{Y}_L$ and $\boldsymbol{H}_L$ are separately fed into the pre-trained VGG19.

III. Experimental Results

This section verifies and compares the proposed method with some representative methods including ITM-CNN [3], FMNet [7], KPN-MFI [9], KUNet [6], SR-ITM [4] and HDR-TV [5], and so on. Moreover, ablation experiment is constructed to investigate the role of the core modules of the proposed method. The proposed method is implemented with Pytorch, and the environment is configured with an Intel(R) Xeon(R) Silver4210 CPU, NVDIA RTX 3090Ti GPU. The proposed FMSTFnet is trained by the Adam optimizer, with β_1=0.9 and β_2=0.999. The batch size is 7, the initial learning rate is set to 0.0002, and it decays to 0.00001 after 100 epochs. The network parameters are initialized by the MSRA tool. A multi-frame SDRv-to-HDRv dataset is constructed for training and evaluation. 20 HDR10 standard HDR videos with 2160×3840 are collected from YouTube,

TABLE I. THE RESULTS OF THE PROPOSED METHOD COMPARED TO THE EXISTING REPRESENTATIVE METHODS

Methods	PSNR↑	SSIM↑	SR-SIM↑	LPIPS↓	ΔEITP↓	HDR-VDP↑
ITM-CNN [3]	29.96	0.9622	0.9358	12.73	22.354	8.0753
FMNet [7]	35.70	0.9811	0.9367	8.78	9.621	8.1787
KPN-MFI [9]	34.73	0.9645	0.9592	14.85	9.733	8.4039
KUNet [6]	35.72	0.9743	0.9419	9.58	10.458	8.2122
SR-ITM [4]	33.89	0.9782	0.9494	10.15	15.522	8.1667
HDR-TV [5]	37.45	0.9858	0.9650	6.53	8.947	8.6111
Proposed	**38.53**	**0.9880**	**0.9710**	**5.34**	**7.517**	**8.6806**

TABLE II. THE RESULTS OF AVERAGE PSNR, SSIM AND ΔE_{ITP} FOR DIFFERENT MODULES

CGFM	AHM	CLFM	HSTNAM1	ViT	HSTNAM2	PSNR↑	SSIM↑	ΔEITP↓
√						36.51	0.9824	9.730
√	√					37.60	0.9862	8.574
√	√	√				37.60	0.9866	8.647
√	√	√	√			37.70	0.9863	8.434
√	√	√	√	√		37.70	0.9869	8.415
√	√	√	√	√	√	**38.53**	**0.9880**	**7.517**

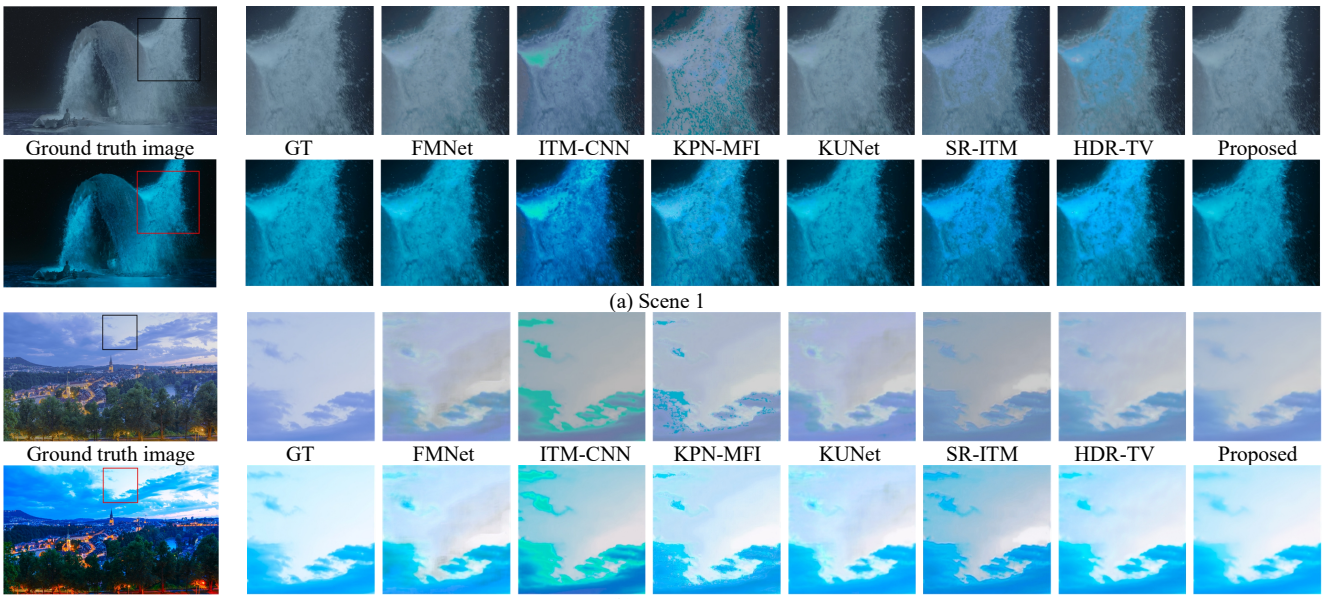


Figure 4. Visual effects of videos obtained by different SDRv-to-HDRv methods (Two partially enlarged regions are water splashes and the sky).

each with a corresponding SDR video. All videos are encoded using PQ curves and BT.2020 color gamut. 16 pairs of videos are used for training, and the remaining 4 pairs are used for testing. To evaluate the quality of generated HDR videos, six quality metrics are used, namely PSNR, SSIM, spectral residual based similarity (SR-SIM), learned perceptual image patch similarity (LPIPS), color difference indicator (ΔE_{ITP}), and HDR visual difference predictor (HDR-VDP).

Table I presents the objective comparison between the proposed method and representative methods, and the best results are presented in bold. The proposed method achieves better HDR video reconstruction performance, resulting in higher fidelity in spatial details and dynamic range of the reconstructed HDR video. The proposed method combines local and global features in the spatial domain and fuses inter-frame features in the temporal domain, this can better fit the nonlinear mapping process required for SDR frame to HDR frame reconstruction. The proposed method also achieves the best performance in ΔE_{ITP}, demonstrating the superiority of the proposed method in color restoration.

Figure 4 shows the visual effects of videos obtained by different methods. For each scene, the upper row shows the original HDR frames without tone mapping, while the lower is the tone mapped frames, similar to [4]. It can be found that the proposed method reconstructs the HDR images with higher visual quality and effectively restores the color information. For example, in the cloud region of the sky, the comparison methods produce significant visual artifacts. In contrast, the proposed method utilizes both local and global information to enhance the reconstruction results, thus more realistically reproducing the information of cloud region.

For the ablation experiments, Table II shows the results of average PSNR, SSIM and ΔE_{ITP} for different modules and their combination. Clearly, the proposed full network

achieves the best performance, which verifies the effectiveness of each module.

IV. CONCLUSIONS

We have proposed a new HDR video reconstruction method from SDR video method based on the design of Feature-Modulation Spatio-Temporal Fusion network (called FMSTFnet). The proposed method can fully utilize temporal and spatial information to reconstruct HDR video, improve the visual effect of the HDR video, and reduce its color deviation. The designed FMSTFnet has low-frequency and high-frequency parts with a pyramid structure, and combined global and local feature modulation module and spatio-temporal fusion module are constructed for eliminating possible inter-frame artifacts and color deviation. In future work, it will be extended to HDR light field reconstruction and angular consistency constraint will be explored to ensure better quality of reconstructed HDR light field images.

ACKNOWLEDGMENT

The work was supported by the National Natural Science Foundation of China under Grant 62271276.

REFERENCES

[1] C. Guo, L. Fan, Z. Xue, and X. Jiang, “Learning a practical SDR-to-HDRTV up-conversion using new dataset and degradation models,” IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 22231–22241.

[2] D. Vo, C. Liu and M. Nelson, “Extremely light-weight learning based LDR to PQ HDR conversion using bernstein curves,” IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2024, pp. 3910–3914

[3] S. Kim, D. Kim, and M. Kim, “ITM-CNN: Learning the inverse tone mapping from low dynamic range video to high dynamic range displays using convolutional neural networks,” Asian Conf. Comput. Vis., 2019, pp. 395–409

[4] S. Kim, J. Oh, and M. Kim, “Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4k UHD HDR applications,” IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 3116–3125.

[5] X. Chen, Z. Zhang, J. Ren, L. Tian, Y. Qiao, and C. Dong, “A new journey from SDRTV to HDRTV,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 4500–4509.

[6] H. Wang, M. Ye, X. Zhu, S. Li, C. Zhu, and X. Li, “KUNet: Imaging knowledge-inspired single HDR image reconstruction,” Int. Joint Conf. Artif. Intell. Eur. Conf. Artif. Intell., 2022, pp. 1408–1414.

[7] G. Xu, Q. Hou, L. Zhang, and M. Cheng, “FMNet: Frequency-aware modulation network for SDR-to-HDR translation,” ACM Int. Conf. Multimedia, 2022, pp. 6425–6435.

[8] L. Xue, T. Xu, Y. Song, Y. Liu, L. Zhang, X. Zhen, and J. Xu, “Lightweight improved residual network for efficient inverse tone mapping,” arXiv preprint arXiv:2307.03998, 2023.

[9] G. Cao, F. Zhou, H. Yan, A. Wang, and L. Fan, "KPN-MFI: A kernel prediction network with multi-frame interaction for video inverse tone mapping", Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, 2022, pp. 806–812.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[11] N. Kitaev, L. Kaiser, and A. Levskaya, “Reformer: The efficient transformer,” arXiv preprint arXiv: 2001.04451, 2020.

InterGridNet: An Electric Network Frequency Approach for Audio Source Location Classification Using Convolutional Neural Networks

Christos Korgialas*, Ioannis Tsingalis*, Georgios Tzolopoulos*, and Constantine Kotropoulos
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, Greece
email: {ckorgial, tsingalis, gtzolopo, costas}@csd.auth.gr
* Equal contribution

***Abstract*—A novel framework, called InterGridNet, is introduced, leveraging a shallow RawNet model for geolocation classification of Electric Network Frequency (ENF) signatures in the SP Cup 2016 dataset. During data preparation, recordings are sorted into audio and power groups based on inherent characteristics, further divided into 50 Hz and 60 Hz groups via spectrogram analysis. Residual blocks within the classification model extract frame-level embeddings, aiding decision-making through softmax activation. The topology and the hyperparameters of the shallow RawNet are optimized using a Neural Architecture Search. The overall accuracy of InterGridNet in the test recordings is 92%, indicating its effectiveness against the state-of-the-art methods tested in the SP Cup 2016. These findings underscore InterGridNet's effectiveness in accurately classifying audio recordings from diverse power grids, advancing state-of-the-art geolocation estimation methods.**

Keywords-electric network frequency (ENF); grid location estimation; audio processing; multimedia forensics

I. INTRODUCTION

Due to power grid disturbances, the Electric Network Frequency (ENF) is a dynamic time series that exhibits fluctuations around its nominal frequency of 50 Hz in Europe and 60 Hz in the United States/Canada. These oscillations result from instantaneous load variations within the power grid, displaying a consistent pattern within interconnected grids. ENF signals become embedded in multimedia recordings captured in proximity to power sources. This distinctive signal can subsequently be extracted [1]–[4] from digital recordings for various applications, such as verification of recording timestamps [5]–[8].

Another application where ENF is also utilized is grid localization. Grid localization can be treated as inter-grid [9]–[11] or intra-grid [10] [12] [13] localization. Inter-grid localization focuses on identifying the grid in which a media recording was captured, while intra-grid localization aims to determine the recording's location within the specific grid precisely. Inter-grid localization is briefly surveyed in Section II.

The intra-grid localization is considered more challenging due to the highly subtle distinctions in the ENF signatures encoded within the recorded signals. However, this assumption is challenged by [14], who detail noticeable differences due to varying city power consumptions and the time for load changes to impact different grid segments, a concept further explored in [15]. Additionally, ENF fluctuations can stem from system disruptions like short circuits, power line switching, and generator failures, as noted in [16]. Minor local load changes affect ENF differently than major events like generator disconnections, which impact the entire grid at about 500 miles per second [17]. Given the aforementioned intra-grid characteristics, various methods have been proposed to tackle the problem of intra-grid localization [18] [19] [20].

A novel framework, termed InterGridNet, is introduced for geolocation classification exploiting the ENF. The framework offers a comprehensive approach that includes data preparation and preprocessing techniques using a shallow RawNet [21] for classification. The topology and the hyperparameters of InterGridNet are optimized through Neural Architecture Search (NAS), enhancing its capability to tackle inter-grid localization in audio recordings. It incorporates innovative techniques, including filtering to isolate the relevant ENF signal, using residual layers to extract frame-level embeddings, and employing a softmax activation function in the decision-making process. To our knowledge, this represents a pioneering advancement spanning from preprocessing techniques to the classification stage, establishing a novel framework in geolocation classification using deep learning methodologies. The Signal Processing (SP) Cup 2016 dataset [22], the only benchmark dataset publicly available, is employed for assessing geolocation classification.

The key contributions of the paper are as follows:

- A novel framework, coined InterGridNet, is proposed to treat geolocation estimation as a classification problem among nine power grids, employing a shallow RawNet optimized with NAS and leveraging ENF signatures from the benchmark SP Cup 2016 dataset. It should be noted that a shallow RawNet is utilized to reduce the number of parameters and achieve comparable performance with that using a deeper neural network.
- Experimental evaluation, including extensive testing of the SP Cup 2016 dataset, showcases the effectiveness of InterGridNet in geolocation classification across nine distinct power grids, where it is compared with state-of-the-art methods.

The remainder of the paper is organized as follows. Sec-

 ISBN: 978-1-68558-245-6

tion II provides an overview of related work. The proposed framework is detailed in Section III. Experimental evaluation is conducted in Section IV. Section V concludes the paper by providing information for future work.

II. Related Work

ENF variations due to load fluctuations and grid frequency control help to localize audio recordings. Grigoras's research demonstrated this by correlating ENF from audio recordings with reference ENF signals from different power grids to estimate the location of the recording [23]. Extensive research was conducted in grid localization using ENF by employing diverse datasets [12]. Additionally, location estimation at various scales was addressed in [24] and [13]. In [10], a machine learning system was developed to ascertain where an ENF-containing media file was recorded, even when no simultaneous ENF reference was available. Five machine learning algorithms were explored to identify the recording location of power and audio recordings obtained from ten distinct power grids in [25]. The hypothesis that variations in load conditions could generate unique location-specific patterns within the ENF signal was assessed in [14]. In [26], an ENF region classification model, UniTS-SinSpec, was introduced within the UniTS framework, integrating a sinusoidal activation function and a spectral attention mechanism and trained on a public dataset. Addressing the complexities of inter-grid classification, field specialists have formulated methodologies to distinguish audio recordings across global power grids, exemplified by the 2016 SP Cup. This work substantially improved the forensic analysis based on ENF, fortifying the verification of the authenticity of multimedia recordings. These distinctive patterns could pinpoint the precise location within a grid where the recording was made.

III. The InterGridNet Framework

A. Dataset Preperation

The SP Cup 2016 competition [22] benchmark dataset [27] is employed, with data split into three sets: a training set for the model's development and training, a practice set for validation, and a testing set for evaluating performance on unseen data (see Section IV-A). The dataset encompasses audio recordings capturing ENF signals from power grids across different countries. Specifically, it consists of recordings from nine distinct power grids, denoted as **A** through **I**. Grids **A**, **C**, and **I** are characterized by nominal ENF at 60Hz, while the remaining grids exhibit ENF around 50Hz.

The dataset consists of audio and power recordings for each grid. The power recordings were generated from a specialized circuit designed to capture the ENF time series directly from the power mains and have a temporal span of 30 to 60 minutes. The audio recordings were acquired using a microphone near substantial electrical devices, capturing the ENF hum for 30 minutes. In particular, power recordings are distinguished by stronger ENF traces than audio recordings.

The testing set has been augmented with 100 samples (40 Audio and 60 Power), each spanning 10 minutes. This subset comprises 8-11 samples from each of the nine grids (**A** - **I**) and 10 additional samples from networks other than these, categorized as "None" (**N**). This diverse sample set is a benchmark for assessing the proposed InterGridNet's efficiency and generalization.

Figure 1 illustrates the InterGridNet framework, highlighting the two critical stages of data preparation and classification. The data preparation process is depicted within the yellow dashed box in Figure 1. Initially, the recordings' inherent characteristics, encompassing ENF signals at either 50Hz or 60Hz, are utilized to classify the recordings as audio or power recordings. Four distinct and independent data groups were created: `audio50`, `audio60`, `power50`, and `power60`.

During the training phase, this categorization is direct since the differences between audio and power recordings are perceptible, mainly due to the higher Signal-to-Noise Ratio (SNR) in power recordings. In contrast, during the testing phase, an automated grouping method is required to classify recordings based on their spectral characteristics, mainly to distinguish between the ENF frequencies of 50Hz and 60Hz. This method is described as follows:

1) For each recording, the average spectrogram magnitude is calculated for the first three harmonics associated with the nominal frequencies of 50 Hz and 60 Hz.
2) For each nominal ENF, the harmonic with the smallest value from step 1 is ignored. Since the ENF may not be present in every harmonic, the two harmonics with the stronger traces are enough for the categorization.
3) The average of the magnitude values at the retained frequencies in step 2 is calculated.
4) The largest value reveals the nominal frequency of the network.

After classifying each recording into its data group, a filtering process is applied to isolate the corresponding ENF within a range of 2 Hz. For instance, samples from the `audio60` group undergo filtering using a bandpass filter set to frequencies between 59 Hz and 61 Hz. Subsequently, the waveforms are segmented into 16-second frames with a 50% overlap and normalized to the interval $[-1, 1]$. These processed frames are subsequently fed into the classification model for power grid classification, shown as the blue dashed box in Figure 1. The aggregated count of frames for each grid is depicted in Figure 2, providing an overview of the distribution of frames across the dataset.

B. Classification Architecture

The spectral content of the frames exhibits variation based on the grid of origin, providing valuable information for the location estimation of the recording. Figure 3 displays a spectrogram concentrated around the nominal ENF from four distinct grids. Notably, the ENF behavior differs depending on the grid, wherein Figures 3(a), 3(b), 3(c), and 3(d) the frequency content is around 60Hz, 50Hz, 50Hz, and 60Hz, respectively. Consequently, the technique elucidated following harnesses these ENF characteristics to classify the samples according to the grid where the recording was made.

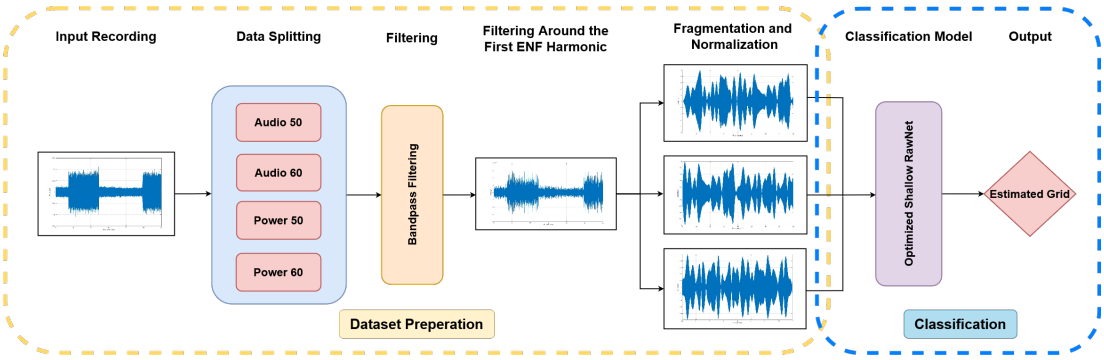


Figure 1. Flowchart of the proposed InterGridNet framework.

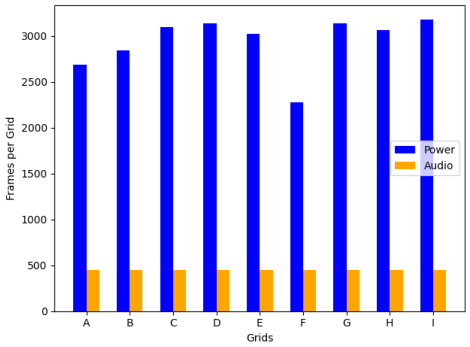


Figure 2. Number of audio and power recording frames in each grid.

TABLE I. OPTIMIZED HYPERPARAMETERS OF THE SHALLOW RAWNET.

	G^{50}_{Audio}	G^{60}_{Audio}	G^{50}_{Power}	G^{60}_{Power}
Learning Rate	6.5×10^{-4}	7×10^{-4}	1.1×10^{-3}	9.7×10^{-4}
β_1	0.96	0.97	0.98	0.98
β_2	0.998	0.998	0.992	0.993

To address the classification problem, individual classes are defined for each grid, comprising 16-second frames derived in Section III-A. These 16-second frames are called samples hereafter. The classification problem for each data group is denoted as $G^{\text{ENF}}_{\text{REC}}$, where REC represents the recording type (Audio or Power), and ENF signifies the nominal frequency of the grid. Consequently, the classification problems are denoted as G^{50}_{Audio}, G^{50}_{Power}, G^{60}_{Audio}, and G^{60}_{Power}. Each $G^{\text{ENF}}_{\text{REC}}$ is expressed as $G^{\text{ENF}}_{\text{REC}} = \{C_1, C_2, \ldots, C_n\}$, where $n = 3$ for G^{60}_{Audio} and G^{60}_{Power}, and $n = 6$ for the others. Each C_i class in the classification problem contains all samples from the corresponding grid in the respective data group.

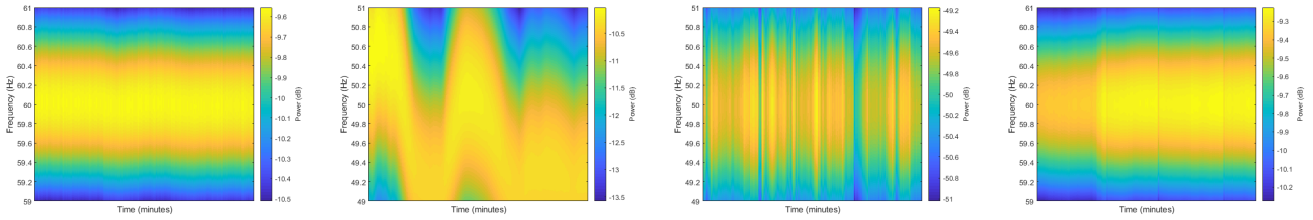
As an illustrative example, the classification problem for the data group `audio60` is denoted as $G^{60}_{\text{Audio}} = \{C_1, C_2, C_3\}$, where C_1 encompasses the audio frames from grid **A**, C_2 from grid **C**, and C_3 from grid **I**. Each sample has a label $l \in \{1, 2, \ldots, n\}$.

For the last part of the flowchart in Figure 1, a shallow RawNet architecture has been implemented to perform the classification. The topology of the shallow RawNet was optimized through NAS using the `Keras-Tuner` library. During this search, several parameters were fine-tuned, including the number of convolutional layers (ranging between 3 and 5), the filter sizes (128 to 256), Gated Recurrent Unit (GRU) units (512 to 1024), and dense layer units (64 to 512). After extensive experimentation, the optimal configuration for this architecture was determined to include two residual blocks.

Specifically, as depicted in Figure 4, the network begins with an input layer that processes frames of size 15,999. These frames are passed through a Strided Convolution block consisting of a one-dimensional convolution layer, batch normalization (BN), and LeakyReLU activation (with a slope of 0.01 for negative inputs). This initial block outputs a feature map of size 5333×128. The first residual block follows, comprising two convolutional layers, batch normalization, LeakyReLU activation, and a max-pooling layer, resulting in an output of 593×128. Following a similar structure, another residual block with 256 filters is applied next, reducing the output to 66×256. These residual blocks are crucial for extracting frame-level embeddings from the input data. Next, the network incorporates a GRU to aggregate these frame-level embeddings into a single ENF-level representation. The output from the GRU is then passed through a dense layer, reducing the dimensionality to a 128-dimensional vector. This layer condenses the extracted features into a more abstract, higher-level representation. Finally, the 128-dimensional vector is processed by the output layer, which uses a softmax activation function to map the vector to a probability distribution over the 9 classes, completing the classification task.

In addition to optimizing the topology of the shallow RawNet, NAS is also employed to fine-tune the hyperparameters. The optimization process explicitly targets the learning rate and parameters associated with the Adaptive Moment Estimation (Adam) optimizer [28]. Initially, the learning rate is set within a range from 10^{-4} to 10^{-2}, and the β values for the Adam optimizer vary between 0.9 to 0.999 and 0.99 to 0.999, respectively. Following the optimization with the `Keras-Tuner` library, the optimal hyperparameter settings for each data group are summarized in Table I. These configurations effectively balance the influence of past and current gradients, contributing to efficient optimization.

To perform grid localization using InterGridNet, we adhere



(a) Grid A (ENF 60Hz). (b) Grid B (ENF 50Hz). (c) Grid D (ENF 50Hz). (d) Grid I (ENF 60 Hz).

Figure 3. Spectrograms focused on the nominal ENF value for different grids.

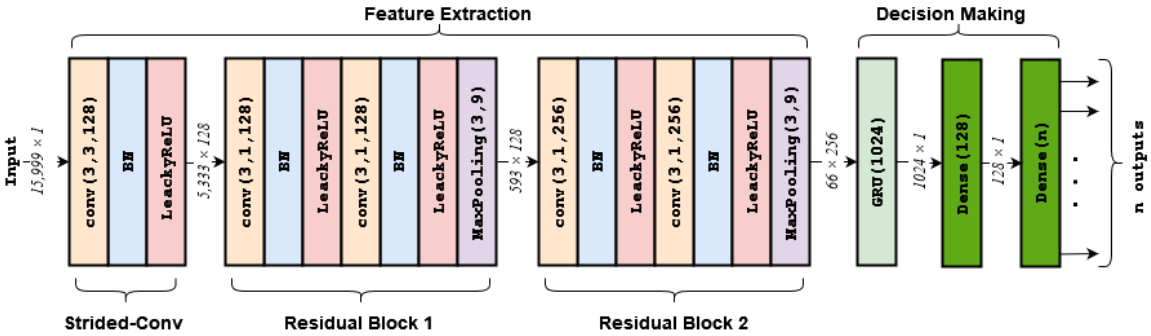


Figure 4. Architecture of the proposed optimized shallow RawNet model. The operators utilized in the network include Conv1D(kernel size, strides, filters), MaxPool1D(pool size, strides), GRU(units), and Dense(nodes).

to the outlined steps depicted in Figure 1. After data preparation, each recording frame undergoes classification by the neural network, resulting in a probability distribution across classes as determined by the softmax activation function of the last layer. For the classification of a frame into one of the known classes, the predictions should satisfy the rule:

$$-\sum_{i=1}^{n} p_i(x) \log_2 p_i(x) < \alpha_1 \cdot \log_2(n), \quad (1)$$

where p_i is the probability for each class prediction from the softmax and n is the number of classes in the frame's data group. In cases where the frame fails to meet (1), it is classified as **N**.

Subsequently, a majority voting mechanism is employed to ascertain the final estimate. The final estimate is deemed valid only if it appears in at least α_2 of the frames' predictions; otherwise, it is designated as **N**. Through the validation process, thresholds α_1 and α_2 have been set to 0.8 and 0.75, respectively. This approach ensures robustness in the grid localization process by requiring a consistent majority agreement across frames for a conclusive final estimation.

IV. Experimental Results

In this section, the validation and testing of the InterGridNet are disclosed [29]. Additionally, limitations are discussed, providing valuable insights into the model's performance and areas for potential improvement.

TABLE II. INTERGRIDNET VALIDATION ACCURACY.

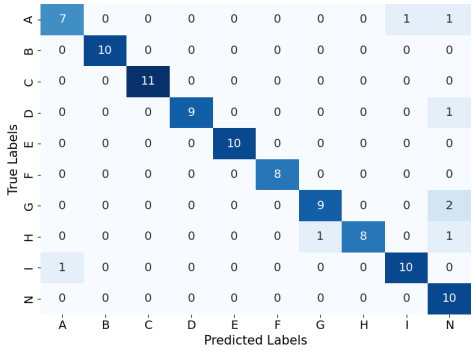
Type	A	B	C	D	E	F	G	H	I	N	Overall
Audio	80%	100%	100%	100%	80%	100%	80%	80%	100%	100%	80%
Power	100%	100%	100%	100%	100%	100%	80%	100%	100%	100%	96.67%
All	80%	100%	100%	100%	80%	100%	60%	80%	100%	100%	90%

A. Model Validation and Testing

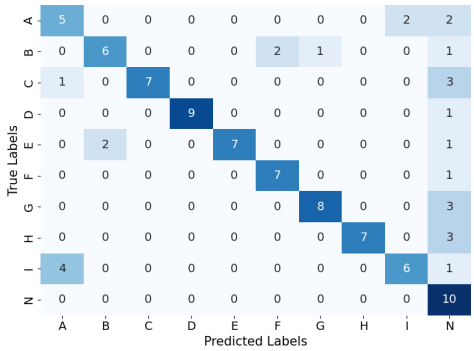
At the training phase of each model, all available training data depicted in Figure 2, corresponding to each data group, were utilized. For validation purposes and to experimentally determine the coefficients α_1 and α_2, the practice set from the SP Cup 2016 dataset was employed. This shares identical characteristics with the testing set described in Section III-A and consists of 50 samples (5 samples for each class).

Table II summarizes the accuracy achieved for each class in the practice set of applying InterGridNet after completing model training and coefficient tuning. The classifier exhibits superior performance in the Power recordings compared to the Audio recordings as the Power recordings contain stronger ENF traces, and the corresponding classifiers benefit from a larger volume of training data, contributing to enhanced performance. In addition, class "None" has 100% accuracy, as shown in column **N**, underscoring the effectiveness of the "None" sample identification method outlined in Section III-B. The aggregate accuracy of the framework culminates at 90%.

The final assessment of InterGridNet's performance was conducted utilizing the dataset testing set. In Figure 5(a), the confusion matrix derived from the predictions of the proposed framework is illustrated, yielding an overall accuracy



(a) Bandpass filtering is applied.



(b) No filtering is applied.

Figure 5. Confusion matrices predictions on the testing set employing the proposed InterGridNet.

of 92%. Notably, misclassifications between classes **A**-**I** are minimal, owing to the inherent constraints of the data splitting technique, which refrains from classifying a sample with ENF at 50Hz into classes **A**, **C**, or **I** with ENF at 60Hz, and vice versa. Consistent with expectations, the testing accuracy closely aligns with the validation accuracy.

B. Discussion

The achieved testing accuracy of 92% underscores the unique characteristics embedded in the ENF signal per grid. Unlike analytical feature extraction methods [30]–[35], these distinctive features, crucial for solving the classification problem, are effectively extracted by the residual blocks and the GRU layer of the neural network described in Section III-B. This observation suggests that the chosen architecture demonstrates exceptional suitability for processing the ENF signal within raw audio data.

Figures 5(a) and 5(b) present the impact of frequency filtering around the nominal ENF on the classification. When this filtering is not applied, the overall accuracy is 72%, significantly lower compared to the scenario with bandpass filtering. This underscores the significant contribution of the ENF signal to accurately determining the grid corresponding to the recording location. In Figure 5(a), the misclassifications by InterGridNet predominantly categorize samples as "None" (class **N**). This exposes a vulnerability of (1) in the framework

TABLE III. TESTING ACCURACIES (%) IN SP CUP 2016 DATASET.

Method	Characteristic	Accuracy
SVM [30]	One-vs-one classification	86%
SVM [31]	Multi-class classification	77%
SVM [35]	Multi-class classification	88%
Random Forest, SVM, AdaBoost [32]	Ensemble method	88%
Binary SVM [33]	Binary classification	87%
Multi-Harmonic Histogram Comparison [34]	Frequency domain analysis	88%
InterGridNet (Ours)	Shallow RawNet	92%

but also underscores its confidence when handling samples from grids on which it has been trained. This dual observation provides insights into the framework's strengths and areas for potential improvement.

Table III summarizes the testing accuracy of other methods using the same testing set. The data highlights the superiority of the proposed InterGridNet framework over previous works, reaffirming its effectiveness in geolocating sound recordings. Hence, InterGridNet is a powerful tool in the field, showcasing its potential for advancing state-of-the-art audio source grid location classification.

In [11], authored by our team, a fusion model comprising five machine learning classifiers was developed, trained, and tested using audio spectrograms from the nine ENF grids. This model achieved a testing accuracy of 96%, compared to the 92% accuracy of the proposed InterGridNet. While the higher accuracy of the fusion model can be attributed to its combination of multiple classifiers, it's important to note that it required a significantly larger parameter count, with 11 million parameters for the CNN alone, which further increased when including the parameters of the fusion framework's classifiers. In contrast, InterGridNet, with a streamlined architecture of 7 million parameters, adopts a novel unified single-classifier approach based on raw audio input via a DNN, highlighting its innovation and efficiency in power grid classification without the need for classifier fusion.

V. Conclusions

This paper presents InterGrid, a novel framework for geolocating audio recordings across different power grids, incorporating optimization through NAS. Inspired by RawNet's architecture, InterGridNet has employed a shallow version of RawNet, offering a dynamic framework that includes preprocessing techniques to tackle the complex challenge of inter-grid localization within audio recordings. Key techniques have been crucial, such as bandpass filtering of ENF data, integration of residual layers for extracting frame-level embeddings, and softmax activation for decision-making. This research has marked the first implementation of DNN methodology for classification with preprocessing methods, achieving a 92% accuracy rate on the SP Cup 2016 dataset. Future research will employ a transformer architecture for grid location classification. To enhance transparency and understand the model's decision-making process, explainable AI (xAI) techniques will also be integrated to extract specific patterns associated with each grid.

ACKNOWLEDGMENT

This research was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the "2nd Call for HFRI Research Projects to support Faculty Members & Researchers" (Project Number: 3888).

REFERENCES

[1] S. Vatansever, A. E. Dirik, and N. Memon, "Analysis of rolling shutter effect on ENF-based video forensics," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 9, pp. 2262–2275, 2019.

[2] C. Korgialas, C. Kotropoulos, and K. N. Plataniotis, "Leveraging electric network frequency estimation for audio authentication," *IEEE Access*, 2024.

[3] C. Korgialas and C. Kotropoulos, "A robust RELAX-based algorithm for enhanced electric network frequency estimation," in *Proc. of the 13th Hellenic Conference on Artificial Intelligence (SETN)*. ACM, 2024, pp. 1–6.

[4] C. Moysiadis, G. Karantaidis, and C. Kotropoulos, "Electric network frequency optical sensing devices," in *Proc. of the 2023 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications*. IARIA, 2023, pp. 1–6.

[5] R. Garg, A. L. Varna, and M. Wu, "Modeling and analysis of electric network frequency signal for timestamp verification," in *Proc. of the IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2012, pp. 67–72.

[6] G. Hua, "Error analysis of forensic ENF matching," in *Proc. of the IEEE International Workshop on Information Forensics and Security (WIFS*, 2018, pp. 1–7.

[7] L. Zheng, Y. Zhang, C. E. Lee, and V. L. L. Thing, "Time-of-recording estimation for audio recordings," *Digital Investigation*, vol. 22, pp. S115–S126, 2017.

[8] G. Hua, J. Goh, and V. L. Thing, "A dynamic matching algorithm for audio timestamp identification using the ENF criterion," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 7, pp. 1045–1055, 2014.

[9] A. Hajj-Ahmad, R. Garg, and M. Wu, "ENF based location classification of sensor recordings," in *Proc. of the IEEE International Workshop on Information Forensics and Security (WIFS)*, 2013, pp. 138–143.

[10] ——, "ENF-based region-of-recording identification for media signals," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 6, pp. 1125–1136, 2015.

[11] G. Tzolopoulos, C. Korgialas, and C. Kotropoulos, "On spectrogram analysis in a multiple classifier fusion framework for power grid classification using electric network frequency," in *Proc. of the 13th International Conference on Pattern Recognition Applications and Methods*. SCITEPRESS, 2024, pp. 91–99.

[12] W. Yao *et al.*, "Source location identification of distribution-level electric network frequency signals at multiple geographic scales," *IEEE Access*, vol. 5, pp. 11 166–11 175, 2017.

[13] R. Garg, A. Hajj-Ahmad, and M. Wu, "Feasibility study on intra-grid location estimation using power ENF signals," 2021, arXiv:2105.00668.

[14] ——, "Geo-location estimation from electrical network frequency signals," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 2862–2866.

[15] M. M. Elmesalawy and M. M. Eissa, "New forensic ENF reference database for media recording authentication based on harmony search technique using GIS and wide area frequency measurements," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 4, pp. 633–644, 2014.

[16] M. H. J. Bollen and I. Y. H. Gu, *Signal Processing of Power Quality Disturbances*. John Wiley & Sons, 2006.

[17] S.-J. Tsai *et al.*, "Frequency sensitivity and electromechanical propagation simulation study in large power systems," *IEEE Transactions on Circuits and Systems*, vol. 54, no. 8, pp. 1819–1828, 2007.

[18] A. Hajj-Ahmad, R. Garg, and M. Wu, "Instantaneous frequency estimation and localization for ENF signals," in *Proc. of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012, pp. 1–10.

[19] Y. Jeon, M. Kim, H. Kim, H. Kim, J. H. Huh, and J. Yoon, "I'm listening to your location! Inferring user location with acoustic side channels," in *Proc. of the World Wide Web Conference*, 2018, pp. 339–348.

[20] Y. Cui, Y. Liu, P. Fuhr, and M. Morales-Rodriguez, "Exploiting spatial signatures of power ENF signal for measurement source authentication," in *Proc. of the IEEE International Symposium on Technologies for Homeland Security*, 2018, pp. 1–6.

[21] J.-W. Jung, H.-S. Heo, J.-H. Kim, H.-J. Shim, and H.-J. Yu, "RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.

[22] M. Wu, A. Hajj-Ahmad, M. Kirchner, Y. Ren, C. Zhang, and P. Campisi, "Location signatures that you don't see: Highlights from the IEEE Signal Processing Cup student competition," *IEEE Signal Processing Magazine*, vol. 33, no. 5, pp. 149–156, 2016.

[23] C. Grigoras, "Digital audio recording analysis–the electric network frequency criterion," *International Journal of Speech Language and the Law*, vol. 12, no. 1, pp. 63–76, 2005.

[24] M. Sarkar, D. Chowdhury, C. Shahnaz, and S. A. Fattah, "Application of electrical network frequency of digital recordings for location-stamp verification," *Applied Sciences*, vol. 9, no. 15, p. 3135, 2019.

[25] Ž. Šarić, A. Žunić, T. Zrnić, M. Knežević, D. Despotović, and T. Delić, "Improving location of recording classification using electric network frequency (ENF) analysis," in *Proc. of the IEEE International Symposium on Intelligent Systems and Informatics*, 2016, pp. 51–56.

[26] Y. Li, T. Lu, G. Zeng, K. Zhao, and S. Peng, "Advanced enf region classification using units-sinspec: A novel approach integrating sinusoidal activation function and spectral attention," *Applied Sciences*, vol. 14, no. 19, p. 9081, 2024.

[27] A. Hajj-Ahmad, "ENF power frequency data for location forensics," https://dx.doi.org/10.21227/H2159S, August 2016, Signal Processing Cup.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] "InterGridNet: An electric network frequency approach for audio source location classification using convolutional neural networks," [Accessed: October 10, 2024]. [Online]. Available: https://github.com/ckorgial/InterGridNet

[30] A. Triantafyllopoulos *et al.*, "Exploring power signatures for location forensics of media recordings," University of Patras, Greece, Tech. Rep., 2016, Signal Processing Cup.

[31] R. Ohib, S. Y. Arnob, R. Arefin, M. Amin, and T. Reza, "ENF based grid classification system: Identifying the region of origin of digital recordings," *Criterion*, vol. 3, no. 4, p. 5, 2017.

[32] M. El Helou, A. W. Turkmani, R. Chanouha, and S. Charbaji, "A novel ENF extraction approach for region-of-recording identification of media recordings," *Forensic Science International*, vol. 155, no. 2-3, p. 165, 2005.

[33] D. Despotović *et al.*, "Exploring power signatures for location forensics of media recordings," University of Novi Sad, Serbia, Tech. Rep., 2016, Signal Processing Cup.

[34] C. Chow *et al.*, "Multi-harmonic histogram comparison," Purdue University, Tech. Rep., 2016, Signal Processing Cup.

[35] H. Zhou, *et al.*, "Geographic location estimation from ENF signals with high accuracy," University of Science and Technology of China, Tech. Rep., 2016, Signal Processing Cup.

Camera Calibration and Stereo via a Single Image of a Spherical Mirror

Nissim Barzilay , Ofek Narinsky , and Michael Werman
e-mail: $nissim.barzilay$@mail.huji.ac.il
$ofek.narinsky$@mail.huji.ac.il
$michael.werman$@mail.huji.ac.il

***Abstract*—This paper proposes a technique for camera calibration and depth estimation from a single view that incorporates a spherical mirror. By leveraging the sphere's contour and reflections, the approach enables precise calibration and scene measurement while capturing additional environmental details beyond the direct image frame. The study explores the geometry and calibration of catadioptric stereo systems, addressing both challenges and practical applications. The paper delves into the intricacies of the geometry and calibration procedures involved in catadioptric stereo utilizing a spherical mirror. Experimental results with synthetic and real-world data demonstrate the method's feasibility and accuracy.**

***Keywords*-Camera matrix calibration; Single-view image; Spherical objects; Mirror sphere; Computer vision.**

I. Introduction

Incorporating spherical mirrors in a catadioptric imaging system makes it possible to observe a wide area with a relatively small mirror. Research and analysis of catadioptric systems based on spherical mirrors can be found in various papers [1]–[3].

Inspired by the concepts introduced in [4], [5], which utilized two spheres in the camera's field of view for obtaining stereo information, our focus is on the more practical scenario of employing a single mirrored sphere. Our research aims to present a method capitalizing on the unique attributes of a single mirrored sphere for both camera matrix calibration and catadioptric stereo.

Our approach only requires the image to show part of the sphere's contour and one of the following; the reflection of the camera, two pairs of corresponding points on and off the spherical mirror, or a single correspondence in special cases.

This research extends to the practical implementation of a real-time system, showcasing the feasibility and efficacy of employing mirrors for stereo imaging as a compelling alternative to the established two-camera stereo methodologies. It is also applicable in scenarios where an accidental spherical mirror is present in the scene. In Section 2, reviews related work in catadioptric imaging system and existing calibration methods, highlighting the advantages and limitations of prior approaches. In Section 3, presents our proposed method for camera calibration and depth estimation using a single spherical mirror, including a detailed explanation of the mathematical formulation and implementation. In Section 4, provides experimental results, including synthetic and real-world data, to validate the accuracy and feasibility of our approach. Finally, In Section 5, discusses the implications of our findings, possible improvements, and potential real-world applications. It concludes the paper with a summary of our contributions and directions for future research.

Figure 1. Spherical mirror in scene

II. Related Work

Catadioptric imaging systems, combining cameras with one or more mirrors, can be divided into categories based on the mirror type and calibration methods. A planar mirror, often used to create a new viewpoint, serves as a cost-effective option for building a stereo system with a single camera. In contrast, a spherical mirror provides a significantly wider field of view, making it popular in catadioptric systems that aim to capture a more complete environment.

Central catadioptric camera calibration: Central catadioptric cameras are imaging devices that use mirrors to enhance the field of view while preserving a single effective viewpoint[6]. Linear calibration methods are proposed that unify the handling of straight-line projections in the real world and sphere images formed by reflections of a spherical mirror, requiring three images of a spherical mirror for implementation. Ying et al. propose a calibration method for paracatadioptric camera based on sphere images, which only requires that the projected contour of a parabolic mirror is visible on the image plane in one view [7]. Their approach relies on the projection properties of spheres in central catadioptric cameras, utilizing a unit viewing sphere model where a sphere projects to two parallel circles they derive constraints for camera calibration. Our method is not sufficient for a central catadioptric camera calibration due to our assumption that the sphere projects an ellipse in the image.

 ISBN: 978-1-68558-245-6

Multiple views of spheres: Agrawal et al. [8], [9] and Zhang [10] developed comprehensive methods for camera calibration, positioning three or more spheres at multiple locations. They present an algorithm that uses the projection of the occluding contours of three spheres and solves for the intrinsic parameters and the locations of the spheres. Extrinsic calibration here involves first estimating each sphere's 3D position in the camera's coordinate system, using known intrinsic parameters and projected ellipses. The methods then determine relative rotation and translation between cameras by aligning these 3D sphere centers. Schnieders et al. [11] propose a method that given multiple views of a single sphere, estimate the camera parameters using the recovered sphere and light directions.

Mirror-Based Calibration with a single-view image: Calibration algorithms that do not require direct observation of 3D reference objects. Many approaches leverage Zhang's calibration algorithm to estimate intrinsic parameters, For instance, Francken [12] utilized this approach for webcam calibration restricted to a screen setup, and others, like Agrawal [13], adapted similar methods.

Perhaps the closest work to our topic is presented by Han et al. [14], who propose a novel self-calibration method for single-view 3D reconstruction using a mirror sphere. Han's approach requires estimating/guessing both the principal point and focal length from a single-view image by minimizing focal length discrepancies between images or through iterative sampling. In contrast, our method computes camera intrinsic parameters directly based on precise mathematical equations derived from the sphere's contour and reflection properties. This approach enables a robust calibration process that avoids iterative estimation, making it suitable for real-time applications.

III. METHOD

In this paper, we assume:

- A projective camera with no skew.
- The image contains a spherical mirror.
- The extrinsic parameters of the camera are

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (1)$$

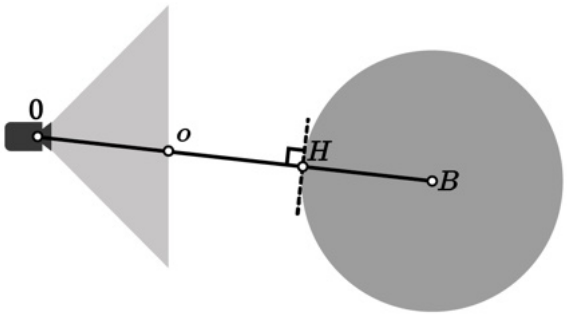
- The unit is defined by the sphere's radius.

To calibrate the camera we need to find the sphere's contour and center in the image. The sphere projects to an ellipse in the image [15]. Let the conic be $v^T C v = 0$, where T denotes transposition, with v the homogeneous coordinates of a point on the conic, and C is the 3×3 symmetric matrix (as illustrated in Figure 1, where the ellipse mark in red represents the projected contour of the sphere.).

Locating ellipses in images is a long-studied challenge, with various methods proposed to tackle it, including both traditional and deep learning approaches, for example, [16]–[19].

Next, we find the sphere's center in the image ($O = \begin{bmatrix} o_x & o_y & 1 \end{bmatrix}^T$). O can be determined by either of the following three methods:

1) Locating the camera's reflection in the mirror (see Figure 9a, Figure 9b). The rays from the camera to the mirror, from the mirror to the camera and the normal at the mirror coincide, thus the ray from the camera to its reflection in the mirror intersects the center of the sphere. So, the image of the camera center is also the location of the *image* of the sphere's center.

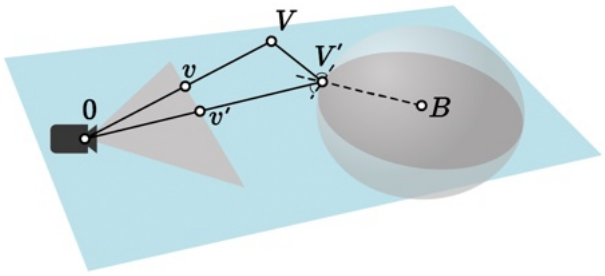


(a) The rays from the camera to the mirror $0 \rightarrow H$, from the mirror to the camera $H \rightarrow 0$, and the normal at the mirror coincide.

(b) The center of the sphere in the image is at the camera's reflection.

Figure 2. Illustration of method 1.

2) Using 2 or more pairs of correspondence points (see Figure 3a, Figure 3b). Let v be the image of a 3D point V and v' the image of V's reflection at V' then the rays from the camera to V', from V' to V and from V' to the sphere's center B (the normal) are coplanar and include the camera center thus project to the line in the image coincident to the sphere's center. The intersection of lines spanning corresponding points, on and off the mirror, is thus the image of the sphere's center.
3) If we assume that the camera has equal focal lengths, $f_x = f_y$. intersecting the line containing a single pair of corresponding points and the major axis of the ellipse, (see Figure 4) suffices. This follows from the *axial* constraint [15], which is the observation that the camera



(a) A 2D cross-section of a pair of correspondence points.

(b) Finding the sphere's center in the image from two pairs of corresponding points.

Figure 3. Illustration of method 2.

Figure 4. Finding the sphere's center from a single pair of corresponding points and the major axis of the ellipse. The green line connects the corresponding points, while the red line represents the major axis of the ellipse.

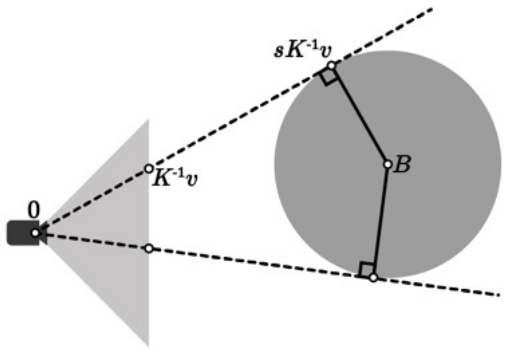


Figure 5. 2D example of sphere outline. $sK^{-1}V$ is perpendicular to $B - sK^{-1}V$. The distance between $sK^{-1}V$ and B is the radius, which is 1.

center, the sphere center, and the major ellipse axis are co-planar. Thus, the image of the sphere center is on the ellipse's major axis.

We want to compute the camera matrix $P_{3\times4}$ and the sphere's center $B = \begin{bmatrix} b_x & b_y & b_z \end{bmatrix}^T$. We will use the radius of the sphere as the unit. Assuming a no skew camera

$$P := \begin{bmatrix} f_x & 0 & t_x & 0 \\ 0 & f_y & t_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{K} & \begin{matrix} 0 \\ 0 \\ 0 \end{matrix} \end{bmatrix} \tag{2}$$

K contains the first 3 columns from the matrix P. Where f_x, f_y are the focal lengths and (t_x, t_y) is the principle point.

Let $V = \begin{bmatrix} v_x & v_y & 1 \end{bmatrix}^T \in \mathbb{R}^3$ be a pixel on the projected contour of the sphere. Geometrically (see Figure 5) this means that there is $s \in \mathbb{R}^+$ such that:

- $\triangle 0, sK^{-1}V, B$ is a right triangle.
  In other words
  $$\langle sK^{-1}V, B - sK^{-1}v \rangle = 0. \tag{3}$$
- The distance between $sK^{-1}V$ and B is the radius.
  The radius is our unit, so
  $$|sK^{-1}V - B| = 1 \tag{4}$$

We simplify these equations to get:

$$\langle K^{-1}V, B \rangle^2 + (1 - |B|^2)|K^{-1}V|^2 = 0 \tag{5}$$

We use the fact that an inner product can be represented by a matrix multiplication and rewrite it as:

$$V^T K^{-T}(BB^T + (1 - |B|^2)I)K^{-1}V = 0 \tag{6}$$

Where I denotes the identity matrix ensuring that it preserves the dimensional of B. This is an equation of the conic section we already computed: C. Therefore, they are equivalent up to a scalar factor:

$$C = rK^{-T}(BB^T + (1 - |B|^2)I)K^{-1} \quad (r \in \mathbb{R}) \tag{7}$$

We currently have 8 unknowns:

$$r, b_x, b_y, b_z, f_x, f_y, t_x, t_y$$

but equating the conic sections only gives 6 equations (Both matrices are symmetric). We first get rid of t_x, t_y by shifting the image so $(0, 0)$ represents the center of the sphere.
We define:

$$S := \begin{bmatrix} 1 & 0 & o_x \\ 0 & 1 & o_y \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

Since we know C, o we can compute the matrix

$$M := S^T C S \tag{9}$$

$$Q := b_z K^{-1} S = \begin{bmatrix} b_z f_x^{-1} & 0 & b_x \\ 0 & b_z f_y^{-1} & b_y \\ 0 & 0 & b_z \end{bmatrix} \tag{10}$$

$$p = \frac{r}{b_z^2} \tag{11}$$

We get:

$$M = pQ^T(BB^T + (1 - |B|^2)I)Q \tag{12}$$

Denote $m_{ij} := M[i, j]$. can be expanded to a system of equations:

$$\begin{cases} m_{11} = p f_x^{-2} b_z^2 (b_x^2 + 1 - |B|^2) \\ m_{22} = p f_y^{-2} b_z^2 (b_y^2 + 1 - |B|^2) \\ m_{33} = p|B|^2 \\ m_{12} = p f_x^{-1} f_y^{-1} b_x b_y b_z^2 \\ m_{13} = p f_x^{-1} b_x b_z \\ m_{23} = p f_y^{-1} b_y b_z \end{cases} \tag{13}$$

To solve these equations, first calculate p and $|B|^2$:

$$p = \frac{m_{13} m_{23}}{m_{12}} \tag{14}$$

,

$$|B|^2 = \frac{m_{33}}{p} \tag{15}$$

Now we can calculate b_x^2, b_y^2, b_z^2:

$$b_x^2 = \frac{1 - |B|^2}{\frac{m_{11}}{m_{13}^2} p - 1}, \quad b_y^2 = \frac{1 - |B|^2}{\frac{m_{22}}{m_{23}^2} p - 1}, \quad b_z^2 = |B|^2 - b_x^2 - b_y^2 \tag{16}$$

The choice of either the positive or negative square root of b_x^2, b_y^2 doesn't matter and it will be compensated by positive or negative f_x, f_y. However, b_z should be positive as the sphere is in front of the camera. Now we can determine the values of f_x, f_y:

$$f_x = \frac{p b_x b_z}{m_{13}}, \quad f_y = \frac{p b_y b_z}{m_{23}} \tag{17}$$

Notice KB is the position of the sphere's center in the image, so $KB = b_z o$. Therefore, we can determine the values based on our previous calculations:

$$t_x = o_x - f_x \frac{b_x}{b_z}, \quad t_y = o_y - f_y \frac{b_y}{b_z} \tag{18}$$

Note that knowing both the sphere's and camera parameters suffice to reconstruct the $3D$ positions of all pairs of corresponding points by intersecting the corresponding rays.

Figure 6. Synthetic Data 1

IV. RESULTS

We have tested our algorithm on a synthetic image of resolution 2048x2048 generated using Blender (see Figure 6), using only the conic section, the contour of the spherical mirror, and the reflection of the camera to calibrate the image.

First phase: We selected points on the sphere contour and calculated the conic. Second phase: We estimate the center of the sphere in the image by locating the camera's reflection. Now we apply our algorithm to calibrate the image. Figure 7 resolution 1920x1080.

TABLE I. COMPARISON OF REAL VALUES AND OUR ALGORITHM'S RESULT ON 6.

	Parameters	
Ground Truth	$b_x = 3$	$b_y = -4$,
	$b_z = 7$	$f_x = 1024$
	$f_y = 1024$	$t_x = 1024$
	$t_y = 1024$	
Result	$b_x = 3.00$	$b_y = -3.94$
	$b_z = 7.03$	$f_x = 1027.99$
	$f_y = 1032.84$	$t_x = 1024.34$
	$t_y = 1016.94$	
Error Range	Less than 1.5%	

TABLE II. COMPARISON OF REAL VALUES AND OUR ALGORITHM'S RESULT ON 7.

	Parameters	
Ground Truth	$b_x = -1.5$	$b_y = 3$,
	$b_z = 1$	$f_x = 1144$
	$f_y = 1144$	$t_x = 960$
	$t_y = 540$	
Result	$b_x = -1.47$	$b_y = 3.07$
	$b_z = 1$	$f_x = 1179$
	$f_y = 1167$	$t_x = 949$
	$t_y = 535$	
Error Range	Less than 3.1%	

Figure 7. Synthetic Data 2

In the real image 1600x1196 (see Figure 1), the estimated sphere origin is:

$$b_x = -0.76, \quad b_y = 0.13, \; b_z = 5.07$$

$$f_x = -1744, \quad f_y = 1732, \quad t_x = 722, \quad t_y = 583$$

To verify our algorithm, we also computed the length of objects using two pairs of correspondence points (Figure 3a) and a sphere with a radius of 5 cm, in Figure 8. We computed the height of the vase using two pairs of corresponding points. We computed the ray for each point. Let v,v', and u, u' be pairs of correspondence points; we then calculate the rays in 3D space. This conversion involves scaling and translating the pixel coordinates. Next, we compute the 3D point representation where the ray intersects the correspondence point v', denoted as h. According to the equation we presented earlier, (4), we define offset $= h - B = sK^{-1}V - B$ with the condition $|sK^{-1}V - B| = 1$. The reflected vector is

$$\text{reflect} = h - 2 * \langle \text{offset}, h \rangle * \text{offset}.$$

Given the reflected ray and the direct ray, we compute the 3D position of the point. The first and second phases are the same as described in the previous example.

$$b_x = 1.30, \; b_y = 0.48, \; b_z = 5.62$$

$$f_x = 2714, \; f_y = 2703 \quad , \quad t_x = 3052, \; t_y = 1664$$

The height of the marker is 13cm, computing the 3D points of v, u marked in red and their distance we obtained is a height of 14 cm. The real height of the tape dispenser is 5cm, computing the 3D points of v, u marked in blue and the distance we calculated a height of 5.05 cm.

TABLE III. COMPARISON OF ZHANG EVALUATION FOR OVER MORE THEN 20 IMAGES AND OUR ALGORITHM'S ON A SINGLE IMAGE RESULT

		Parameters			
9.	Zhang Calibration	$f_x = 8146$			
		$f_y = 8286$	$t_x = 3143$	$t_y = 2397$	
	Result	$f_x = 8258$			
		$f_y = 8073$	$t_x = 3904$	$t_y = 3875$	.

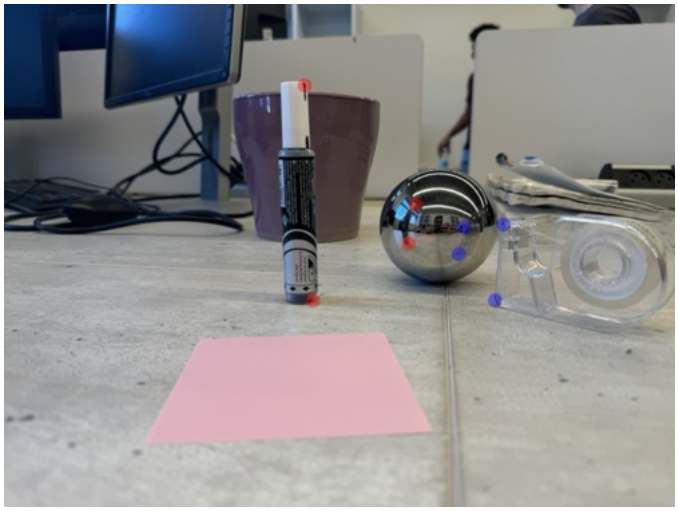


Figure 8. Height test

(a) Single image of a spherical mirror - our algorithm

(b) Images - Zhang algorithm

Figure 9. Comparison of Camera Calibration Methods for the Canon EOS R10

V. CONCLUSION AND FUTURE WORK

We presented a novel approach for calibrating the camera matrix using a single-view image. Our findings help reduce the requirements for achieving this calibration. Using our method, further image analysis is possible, such as determining the 3D location of a point from a pair of corresponding points or estimating an omnidirectional image centered at the sphere's origin. Additionally, since a spherical mirror distorts the scene by projecting it onto a curved surface, we aim to leverage our findings to correct this distortion and reconstruct the scene as if it were reflected in a planar mirror in future work.

REFERENCES

[1] S. Barone, M. Carulli, P. Neri, A. Paoli, and A. V. Razionale, "An omnidirectional vision sensor based on a spherical mirror catadioptric system," *Sensors*, vol. 18, no. 2, 2018. DOI: 10.3390/s18020408.

[2] Y. Hiruta, C. Xie, H. Shishido, and I. Kitahara, "Catadioptric stereo-vision system using a spherical mirror," *Procedia Structural Integrity*, vol. 8, pp. 83–91, 2018, AIAS2017 - 46th Conference on Stress Analysis and Mechanical Engineering Design, 6-9 September 2017, Pisa, Italy, ISSN: 2452-3216. DOI: https://doi.org/10.1016/j.prostr.2017.12.010.

[3] Y. Hang, L. Zhao, and W. Hu, "A survey of catadioptric omnidirectional camera calibration.," *International Journal of Information Technology and Computer Science,*, 2013.

[4] S. Nene and S. Nayar, "Stereo with mirrors," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 1998, pp. 1087–1094. DOI: 10.1109/ICCV.1998.710852.

[5] Y. Hiruta, H. Shishido, and I. Kitahara, "One shot 3d reconstruction by observing multiple spherical mirrors," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 2021, pp. 580–584. DOI: 10.1109/GCCE53005.2021.9621989.

[6] X. Ying and H. Zha, "Identical projective geometric properties of central catadioptric line images and sphere images with applications to calibration," *International Journal of Computer Vision*, vol. 78, no. 1, pp. 89–105, Oct. 2007. DOI: 10.1007/s11263-007-0082-8.

[7] Y. Li and Y. Zhao, "Calibration of a paracatadioptric camera by projection imaging of a single sphere," *Appl. Opt.*, vol. 56, no. 8, pp. 2230–2240, Mar. 2017. DOI: 10.1364/AO.56.002230.

[8] Agrawal and Davis, "Camera calibration using spheres: A semi-definite programming approach," in *Proceedings Ninth IEEE International Conference on Computer Vision*, 2003, 782–789 vol.2. DOI: 10.1109/ICCV.2003.1238428.

[9] M. Agrawal and L. Davis, "Complete camera calibration using spheres : A dual-space approach," Jan. 2003.

[10] H. Zhang, K.-y. K. Wong, and G. Zhang, "Camera calibration from images of spheres," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 499–502, 2007. DOI: 10.1109/TPAMI.2007.45.

[11] D. Schnieders and K.-Y. K. Wong, "Camera and light calibration from reflections on a sphere," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1536–1547, 2013, ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2013.06.004.

[12] Y. Francken, C. Hermans, and P. Bekaert, "Screen-camera calibration using a spherical mirror," in *Fourth Canadian Conference on Computer and Robot Vision (CRV '07)*, 2007, pp. 11–20. DOI: 10.1109/CRV.2007.59.

[13] A. Agrawal, "Extrinsic camera calibration without a direct view using spherical mirror," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 2368–2375. DOI: 10.1109/ICCV.2013.294.

[14] K. Han, K.-Y. K. Wong, and X. Tan, "Single view 3d reconstruction under an uncalibrated camera and an unknown mirror sphere," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 408–416. DOI: 10.1109/3DV.2016.50.

[15] T. Tóth and L. Hajder, "A minimal solution for image-based sphere estimation.," *Int J Comput Vis*, 2023.

[16] M. Cicconet, K. Gunsalus, D. Geiger, and M. Werman, "Ellipses from triangles," in *2014 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2014, pp. 3626–3630.

[17] A. Fitzgibbon, M. Pilu, and R. B. Fisher, "Direct least square fitting of ellipses," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 5, pp. 476–480, 1999.

[18] C. Lu, S. Xia, M. Shao, and Y. Fu, "Arc-support line segments revisited: An efficient high-quality ellipse detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 768–781, 2019.

[19] W. Dong, P. Roy, C. Peng, and V. Isler, "Ellipse r-cnn: Learning to infer elliptical object from clustering and occlusion," *IEEE Transactions on Image Processing*, vol. 30, pp. 2193–2206, 2021.

Efficient Implementation of CNN in Deep Learning by Using the Multirate Algorithms

Guowei Xiao
Faculty of Automation Dept, University of Technology, Guangzhou, 510006, China, 2112204008@mail2.gdut.edu.cn

Yingshuai Wang
Chendian Technology (Shanghai) Co., Ltd, tiger.wang@chendiantech.com

Ping Wang
San-xin-dui Tech, 200030, China wang@novasense-tech.com

***Abstract* — This paper proposes the multirate Convolutional Neural Networks (CNN) algorithms for an efficient implementation of the 2-Dimensional (2-D) CNN circuits implementation. During the rapid growth in computation power, Deep Learning (DL) using CNN has widened the areas of the Artificial Intelligent (AI) applications. For the layers of the convolution with pooling operation in CNN some researchers work has initially applied the multirate algorithms to the traditional (non-multirate) convolutional kernel operation of using polyphase architectures resulting in the more efficient implementation of the multirate filtering. In this work we extend it into 2-D CNN by using time-varying coefficient to achieve an efficient implementation with reduced memory(i.e. the line-buffer) size by M-fold(the pooling factor) and the MACs at 1/M of clock running rate. A design example of the first stage of CNN system will be provided. Its results are verified with the Matlab CNN-based digit recognition tool.**

Keywords—CNN; ML; DL; AI; IC; Multirate; 2-D; Signal Processing; DSP; AISC; Filter.

I. INTRODUCTION

With the surging of the computational power, Deep Learning (DL) using Convolutional Neural Networks (CNN) has become reality in more and more applications of Artificial Intelligence (AI). However, some applications have limited energy capacities. In various Internet of Things (IoT), the wearable and mobile applications of CNNs have scarce energy sources and thus require solutions to lower power consumption and smaller hardware size in order to ensure the longevity of the devices and smaller chip area [4]. As the result of the demand for lower power consumption, more research interest has been generated in exploring high-performance neural processing units or Application Specific Integrated Circuit (ASIC) accelerators with superior power efficiency and computation parallelism [5].

Fig. 1 shows a typical DL system with CNN architecture which contains convolution, pooling, and fully-connected layers. It usually includes several cascaded convolutional layers in which the a single clock frequency is employed [5].

In the applications of real-time image processing, for instance, the 2-D CNN hardware architectures that make dense, pixel-wise predictions, such as FCN [6], U-Net [7], and their variants, use very long skip lines. For example each line contains as many as 512 pixels in the U-Net image. Those skip lines are crucial for recovering of the details lost during the down-sampling. The IC hardware implementations of those networks require large memory (or line-buffer or line-delay) to store all the skip lines. The line buffers often use external memory, such as SRAM or DDR, which dramatically increases the cost in terms of silicon area footprint and consumes high power [8].

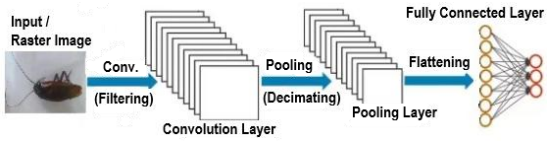


Figure 1. A typical CNN architecture with convolutional, pooling and full-connected layers

Images or 2-D signals are acquired line-by-line by the raster scan sequence. In the 2-D raster scanning-line based system, the row (vertical) delay of lines is accumulated in each convolution layer. For example, a 2x2 spatial window may cause a 1-line delay, whereas two consecutive 3x3 convolutions may result in 2-line delays and each delay contains 512 pixels to be stored in the U-Net image.

The problem with the long skip lines is that once the data on one end of the skip line is generated, it needs to be held in memory until the data in the receiving end of the skip connection is available. The more layers a connection skips over, the more line pixels need to be stored in memory. Therefore, the size of the total memory required increases with the length of the skip line. The memory requirements for the line delays can aggregate quickly and become a significant contributor to the total silicon area needed to implement the network. Moreover, the latency issues can also be problematic in latency-sensitive applications such as autonomous driving systems.

The computation of convolutional operations involves multipliers and adders, i.e., the Multiply-and-Accumulate (MAC) operation. For concurrent processing, the number of multipliers required must be the same as the filter size, which can result in large area consumption. Moreover, summing up the outputs of these multipliers involves multiple cascaded adders. Thus, digital MAC units may occupy a vast area with high power consumption [8].

The chip area and power constraint facilitate the researcher interests in the multirate filtering techniques [2] which can not only perform real-time kernel convolution but can also occupy significantly less chip area and smaller power consumption. Although the works in [2][3] are in the

analog-digital mixed-signal domain, their multirate (decimating filter) algorithms and implementation architectures can be expanded into the digital signal processing domain.

In this paper we describe the way to design decimating (multirate) filters for kernel convolution with pooling (decimating) operations, and introduce the time-varying coefficient (weight)architectures for the efficient 2-D CNN circuit implementation architectures whose memory (the line-buffer) size is to be reduced by M-fold(a pooling factor) and the MACs at 1/M of the clock rate.

The paper is organized as follows. In Section II, the multirate algorithms for 1-D decimating filter is presented in terms of time-varying coefficients. In Section III, a direct-form implementation of the 2-D CNN counterpart is derived. Finally, Section IV presents a design example of 3x3 kernel convolutional layer with pooling 2x2 (decimating filter) for demonstrating of the 2-D CNN implementation.

II. MULTIRATE ALGORITHMS FOR 1-D CONVOLUTION WITH POOLING OPERATIONS IN CNN

The multirate algorithms for efficient implementation of 1-D and 2-D filtering circuits have been previously introduced by [1][2][3] based on polyphase structures. In CNN efficient implementation, however, we modify the polyphase structures and manipulate the (decimating) filter transfer functions as filtering with the time-varying coefficients (weights) form. Thus, the resulting filter expression form is comparable to its non-multirate prototype counterpart.

The multirate algorithms in terms of time-varying coefficient expression give explicit mapping relations between non-multirate and multirate relations of z-transform functions. These can be utilized for efficient design and implementation architectures of such 1-D and 2-D decimating filters.

For the sake of easy comprehension, only the first (one) layer of CNN in Fig. 1 is discussed and illustrated. We can see that the convolutional and pooling layers architecture is the same as the 2-D decimating filtering system [2][3], as depicted in Fig. 2, with an activation operation (ReLU) which operates either after or prior to the pooling.

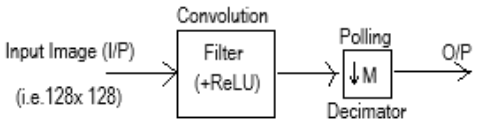


Figure 2. The convolution followed by pooling architecture can be considered as a 2-D decimating filter.

To derive the efficient implementation architecture we further consider a 1-Dimensional (1-D) linear, time-invariant (N-1)-th order FIR filter followed by a decimator with a factor (In neural network computation, stride for pooling layers is often used） of M, as illustrated in Fig. 3(a) below.

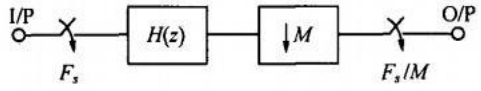


Figure 3. (a) A general filter clocking at F_s and followed by a decimation operator

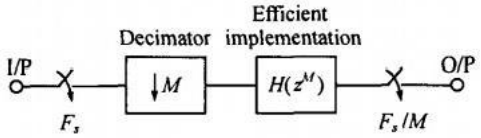


Figure 3. (b) Deriving efficient multirate implementation for an1-Dflter.

and its z-transfer function is $H(z)$ as shown in Eq.1, where the unit-delay z^{-1} is related to the sampling frequency F_s . The overall system clocking at an unique frequency F_s is also called non-multirate (traditional) system,

$$H(z) = \sum_{n=0}^{N-1} h_n z^{-n} \qquad (1)$$

For an efficient implementation of Fig. 3(a) this system $H(z)$ can be alternatively manipulated as Eq.2 by using multirate (decimation or interpolation) filter architecture based on the polyphase decomposition algorithms as described in [1][2].

The efficient implementation implies that the most parts in CNN operate at lower clock frequency (lower power consumption) and less memory used (smaller memory size required) especially in the 2-D and 3-D [9] CNN systems. Such a decimating filter using time-varying coefficients (weights) of convolutional layer expression can be considered as a time-variant filter with periodically varying coefficients [2][3]. It can be straightforwardly applied to CNN implementation in which the convolutional layer is followed by the pooling operation.

Considering such a decimating filtering system as in Fig. 3(a) which can be mathematically expressed as Eq.1, where the (N-1)-th order prototype filter with decimating factor (pooling stride) of M, it can be manipulated as

$$\begin{aligned} H(z) &= h_0 + h_1 z^{-1} + h_2 z^{-2} + \ldots + h_{N-1} z^{-(N-1)} \\ &= \left(h_0 + h_1 z^{-1} + h_2 z^{-2} + \ldots + h_{M-1} z^{-(M-1)}\right) \\ &+ \left(h_M + h_{M+1} z^{-1} + \ldots + h_{2M-1} z^{-(M-1)}\right) z^{-M} + \ldots \\ &+ \left(h_{N-M+1} + \ldots + h_{N-1} z^{-(M-1)}\right) z^{-(L-1)M} \end{aligned} \qquad (2)$$

where the filter order N= ML. Eq.2 contains L terms (and each of which contains bracketed M sequential terms that can be considered as a periodically commuted coefficient. We define such a coefficient as a time-varying one. Therefore, it has L time-varying coefficients.

Assuming Z= (z^M) which is related to the reduced sampling rate F_s/M . Thus, we arrive at the transfer function with a time-varying coefficient form:

$$H(Z) = \tilde{h}_0 + \tilde{h}_1 Z^{-1} + \ldots + \tilde{h}_{L-1} Z^{-(L-1)} = \sum_{i=0}^{L-1} \tilde{h}_i Z^{-i} \quad (3)$$

where h_i represents the time-varying weights in Eq.3. It is noticed that Eq.3 has a similar math expression form to its non-multirate (prototype filter) counterpart. $H(Z)$ or $H(z^M)$ is operating at F_s/M which is a lower clock rate than the original F_s.

III. Efficient Implementation of 2-D Multirate Convolutional and Pooling Layers in CNN

Fig. 4 shows an FIR prototype 2-D filter with the transfer function $H(z_1,z)$ where z^{-1} represents the horizontal-dimensional delay unit and z_1^{-1} represents the vertical-dimensional delay unit(scan-line delay). The overall filter system is operating at the horizontal frequency F_s. This non-multirate 2-D filter can be expressed as Eq.4.

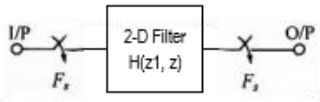


Figure 4. A non-multirate prototype 2-D filter.

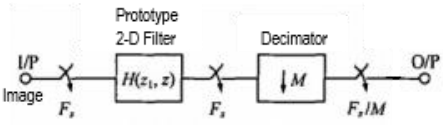
$$H(z_1, z) = \sum_{j=0}^{N-1} \sum_{i=0}^{N_1-1} a_{ij} z_1^{-i} z^{-j} \quad (4)$$

where a_{ij} are the normalized weight coefficients for both the horizontal and vertical dimensions. The index i is equal to the integer of for i = 0,1, and (N_1-1) where N_1 is defined as the filter order in the vertical dimension. Similarly, index j is for 0,1, and (N-1) where N is the horizontal dimension filtering order.

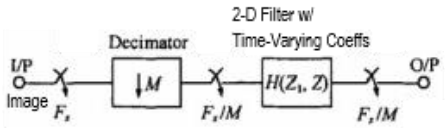
The variable separable filters (convolutions) are commonly used to design efficient neural network architectures [8]. For the demo purpose of multirate concept, assume the $H(z_1,z)$ to be variable separable. Therefore Eq.4 can be further simplified to Eq.5 [2][3],

$$H(z_1, z) = H(z_1)H(z)$$
$$H(z) = \sum_{j=0}^{N-1} a_j z^{-j} \quad (5)$$
where $$H(z_1) = \sum_{i=0}^{N_1-1} a_{1i} z_1^{-i}$$

Assuming that decimating factor M is the same in both dimensions, in Fig. 4 we apply the multirate transformation [3] to both $H(z)$ and $H(z_1)$ as similar form to Eq.3, and thus an efficient implementation can be achieved in which the scan-line memory length and computational clock speed can be reduced by a factor of M as shown in Fig. 5.



(a)A 2-D non-multirate filter $H(z_1,z)$ followed by a decimator



(b)The efficient implementation form of a 2-D decimating filter and the decimator is now in front of filter

Figure 5. Deriving the efficient multirate implementation for the 2-D filter.

Fig. 5(a) and (b) depicts the process of deriving efficient implementation of the 2-D decimating filter. It can be observed in Fig. 5(a) a typical convolutional layer followed by a pooling layer, in which the filter circuit is operating at the system maximum frequency F_s and the scan-line memory is equal to the input image pixel numbers in each line.

In Fig. 5(b), however, the decimator is placed in the front of the 2-D filter and it yields an efficient implementation when using time-varying weights. This can be described as the following Eq.6.

$$H(Z_1, Z) = H(Z_1)H(Z)$$
$$H(Z) = \sum_{j=0}^{L_2-1} \tilde{A}_j Z^{-j} \quad (6)$$
Where $$H(Z_1) = \sum_{i=0}^{L_1-1} \tilde{A}_{1i} Z_1^{-i}$$
$$Z = z^M$$
$$Z_1 = z_1^M$$

where the capital-case z_1^{-1} represents the vertical scan-line delay and the capital Z equals to (z^M), so Z^{-1}= (z^M)$^{-1}$ which implies that the computation rate (the required sampling frequency) has been lowered with the factor of M. L_1 and L_2 are the time-varying coefficient indexes, respectively. Thus, we arrive at an efficient implementation architecture of Eq.6 with the time-varying weights as shown in Fig. 6

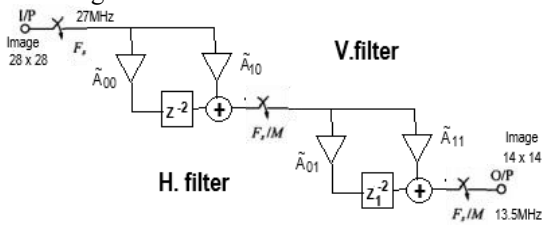


Figure 6. The proposed efficient implementation with time-varying weights

where M equals 2 in both dimensions. Thus, the time-varying coefficients can be manipulated as

$$\tilde{A}_{00} = a_0 + a_1 z^{-1}; \tilde{A}_{10} = a_2 + a_3 z^{-1}; \text{ and}$$

$$\tilde{A}_{01} = a_{10} + a_{11}z_1^{-1}; \tilde{A}_{11} = a_{12} + a_{13}z_1^{-1};$$

IV. A DESIGN EXAMPLE OF 2-D MULTIRATE CNN

Consider a 2-D FIR edge-filter example whose coefficients is listed in TABLE I below

TABLE I. THE 2-D SEPERABLE EDGE FILTER WITH BOTH DIMENSIONAL WEIGHTS

	Filter Coefficients (Weights)
Horizontal Filter $H(z)$	a_0= -3.9; a_1= 0; a_2= 4; a_3= 0;
Vertical Filter $H(z_1)$	a_{10}= -3.9; a_{11}= 0; a_{12}= 4; a_{13}= 0;

For comparison, an image as shown in Fig. 7(a) inputs to the three types of multirate (decimation) filter shown in Fig. 6.

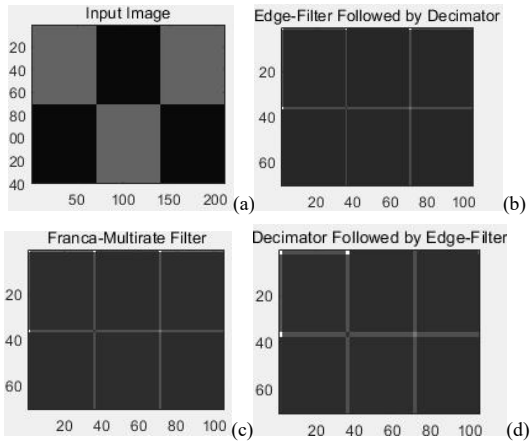


Figure 7. (a)Input image; (b)Type-I:Traditional convolution followed by pooling system's output image; (c)Type-II:Proposed multirate filter output; (d)Type-III:The pooling layer followed by the convolution layer.

The type-I is the same as convolution layer of 3x3 kernel followed by the pooling layer with stride =2 in CNN and the simulated output image is as shown in Fig. 7(b). The type-II is the proposed Franca-multirate edge filter architecture and the output image is as shown in Fig. 7(c); The type-III consists simply of placing the decimator in front of the filter and the output is shown in Fig. 7(d).

Comparing the above mentioned output images, we notice that the proposed Franca-multirate filter has the same output with the traditional convolution plus pooling's output. To further verify the multirate architecture, consider again the case of the design example for a convolution layer with pooling stride=2. The operating clock frequency is set at 27MHz. It can be seen that the entire 2-D filter now operates at a lower frequency 13.5MHz which can reduce power consumption in the circuit and the scan-line memory by half. In addition, the feature-map memory of CNN is also reduced by three quarters (image 14x14).

By using MATLAB CNN based tool at 3x3 for the digit-recognition, we compare the simulation results from the original code to modified code which models our multirate architecture in the first convolution, ReLU, and pooling layers as shown above in Fig. 5 where the bias values have been considered in the weights during the training..

The weights training has no noticeable delay or any convergence issue, and the final detecting accuracy is identical to the MATLAB original results as depicted in Fig. 8.

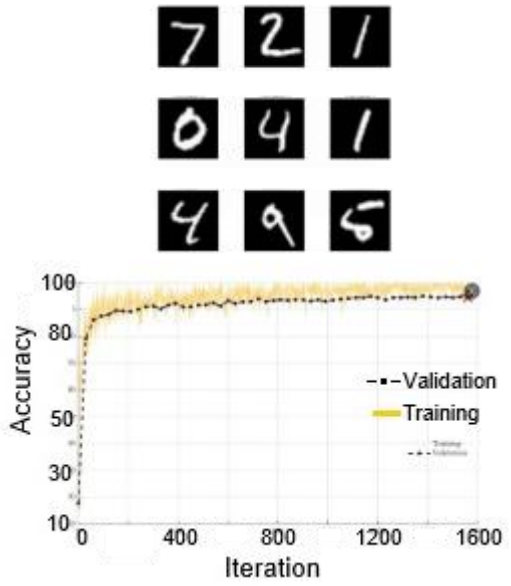


Figure 8. Digit-recognition simulation results of the multirate 2nd-order filter with pooling stride=2

V. CONCLUSIONS

We have proposed the new multirate algorithms with time-varying weight architectures for efficient CNN hardware implementation. The design example has been verified with digit-detection CNN-based MATLAB tool. It has achieved 2-fold reduction of computing clock rate and line delay memories for the CNN implementation resulting in a smaller chip size and lower power consumption.

As future work, it would be interesting to design ASIC chips to study how the efficient implementation of the while multirate CNN presented in this paper would applied into many applications in DL of AI, especially in the 2-D and 3-D CNNs. The training methodology of the multirate CNN should be further studied to achieve a similar generalized existing learning methods.

ACKNOWLEDGEMENT

The authors would like to thank previous colleagues at Instituto Superior Tecnico (IST) -Mr. Paulo Santos and

Shanghai Jiao-Tong University -Dr. Haishan Wu for their kindly reviews.

REFERENCES

[1] R. Crochiere and L. R. Rabiner, "*Multirate Digital Signal Processing",* Englewood Cliffs, NJ: Prentice-Hall, 1983.

[2] J. E. Franca,"*Non-Recursive Polyphase Switched-Capacitor Decimators and Interpolators*", IEEE Trans. Circuits Syst., vol. CAS-32, pp877-887, Sept.1985.

[3] P. Wang and J. E. Franca,"*Multirate Switched-Capacitor Circuits for 2D Signal Processing*", Kluwer Academic Publishers, ISBN 0-7923-8051-7, 1998.

[4] Y. Le Cun, Y. Bengio, and G. Hinton, "*Deep learning,*" Nature, vol. 521, pp. 436–444, May 2015.

[5] S. Gupta, "*Neuromorphic Hardware: Trying to Put Brain into Chips*.", 30 June 2019. Available online:https://towardsdatascience.com/neuromorphic-hardware-trying-to-put-brain-into-chips-222132f7e4de (accessed in 2023).

[6] Q. Chen, J. Xu,, and V. Koltun, "*Fast image processing with fully-convolutional networks*.", In Proceedings of the IEEE International Conference on Computer Vision, pp.2497–2506, 2017.

[7] O. Ronneberger, P. Fischer, and T. Brox, "*U-net: Convolutional networks for biomedical image segmentation*" , In International Conference on Medical Image Computing and Computer-assisted Intervention, pp.234–241. Springer, 2015.

[8] M. Asama, L. Isakdogan, S. Rao, B. Nayak and G. Micheal,"*A Machine Learning Imaging Core Using Separable FIR-IIR Filters*", Xiv:2001.00630v1[eess.IV] Jan 2020, Intel Co

[9] H. A. Madanayake and L. Bruton, "*A Systolic-Array Architecture for First-Order 3-D IIR Frequency-Planar Filters*", IEEE Trans. on CAS-I:, VOL.55, NO. 6, pp.1546-1559, July 2008.

Finite Word Length Effect in Practical Block-Floating-Point FFT

Gil Naveh
Tel-Aviv Research Center
Huawei Technologies Co. Ltd
Tel-Aviv, Israel
Email: Gil.naveh@huawei.com

***Abstract*— Fixed-Point FFT implementation is very sensitive to finite-word-length-effects due to the large quantization noise that is being accumulated throughout the FFT stages. In FFT implementations on fixed register size processors like CPUs and DSPs, Block-Floating-Point (BFP) is a well-known scheme for controlling the tradeoffs between the fixed-point register size and the resultant accuracy. The performance of the ideal BFP FFT, in terms of the output Signal to Quantization Noise Ratio (SQNR), has been investigated in depth. However, ideal BFP-FFT suffers from implementation complexity, and especially non-deterministic latency. This is caused by the inherent mechanism that requires to re-calculate an entire FFT stage if one of the stage's output overflows. Because of this, most of the implementations are of a more practical variant for the BFP-FFT that does guarantee fixed latency. This, however, comes on the expense of reduced accuracy (degraded SQNR). In this paper, we derive the SQNR formulas for the practical BFP-FFT for radix-2 and radix-4 Cooley-Tukey Decimation-In-Time (DIT) FFTs. The derived model is compared to computer simulations and found highly accurate (less than 0.2dB difference). We use the derived model to compare the SQNR performance of the practical algorithm to the ideal one and show a 6-14dB penalty cost for guaranteeing fixed latency implementation.**

Keywords - Block Floating Point; Fixed Point; DIT; SQNR;

I. INTRODUCTION

The Fast Fourier Transform (FFT) serves as an important tool in many signal processing applications. Throughout the Years it has been successfully used in radar application, spectral analysis, filtering, voice enhancement, advanced audio codecs (like MP3 and AAC), and during the last three decades, with the introduction of multitone modulations, it is also being successfully used in wired and wireless modems such as discrete-multi-tone in Digital-Subscriber-Line (DSL) modems [1], Orthogonal-Frequency-Division-Modulation (OFDM) in several wireless modems, e.g., [2] and in advanced fiber optic modems [3].

Finite-word-length effects (denoted hereafter also as quantization noise) have substantial effect on the accuracy performance of FFTs. This is a result of the native characteristic of the FFT in which quantization noise that is added at the output of each stage of the FFT is accumulated toward the FFT output. Since the maximal value at each stage's output grows as we proceed with the stages [4], in many hardware implementations, the performance degradation due to the quantization noise is mitigated by adapting the register size at each stage to accommodate the signal growth [5]-[7]. On the other hand, in software implementations (as in CPUs and Digital Signal Processors - DSPs), or hardware implementations where intermediate values are forced to be written to memory, increasing the bit-width of the stored values is not possible. For those cases, a dynamic-scaling BFP based schemes are commonly used.

The straight-forward dynamic-scale BFP is such that throughout the calculation of each FFT stage, the butterflies' outputs are tested for an overflow. If an overflow is detected, the entire stage is recalculated and scaled down before stored to memory. The advantage of this BFP scheme is that the scale down is done only on a concrete need, which leads to the best accuracy performance among other BFP-FFT schemes. For that reason, we relate to the straight-forward dynamic-scale BFP-FFT as "ideal BFP-FFT" herein. The drawbacks of this scheme are its complexity and the fact that it results in non-deterministic latency. Deterministic latency may have high importance when the FFT is used within a synchronized pipelined system, such as a modulator or demodulator in OFDM modems [8].

Multiple schemes that overcome the non-deterministic latency drawback have been proposed, e.g., [9] [10], but they all involve non-negligible SQNR performance degradation as compared to the ideal BFP. Among the class of the deterministic latency BFP-FFTs, the one proposed by Shively [11] leads to the least SQNR loss as compared to the ideal BFP-FFT. Thanks to this fact, it turns to be among the most common schemes for practical implementations, e.g., [12] [13]. We refer to the Shively's scheme herein as "practical BFP-FFT".

The ideal BFP-FFT was originally analyzed in [14], which provided a lower and upper bound for the output quantization noise variance. In [4] and [9], a more accurate statistical model was used to project the expected value of the ideal BFP-FFT output noise power for an uncorrelated input sequence. Although the practical BFP-FFT is widely used in practical systems for deterministic latency BFP-FFT, to the best knowledge of the author, its accuracy performance has not been analyzed.

In this work, we refine the commonly used statistical model of quantization noise within FFTs, apply this refinement to the SQNR of the ideal BFP-FFT, and derive the analytical model of the SQNR of the practical BFP-FFT. We adapt the noise models to represent modern processor having

 ISBN: 978-1-68558-245-6

embedded complex multipliers and wide accumulators, and we evaluate the accuracy degradation of the practical BFP-FFT as compared to the ideal one.

The paper is organized as follows: Section II introduces the models used throughout the paper covering the DIT FFT model, the underline processor model, and the quantization noise models. In Section III the analytical SQNR formulas of a generic scaling policy are derived and in Section IV the associated scaling policies for the ideal and practical BFP FFT are described. Section V applies the SQNR formulas to the associated scaling policies while the results are presented in Section VI., Finally, conclusions are given in Section VII.

II. FFT, PROCESSOR AND QUANTIZATION NOISE MODELS

We relate to fixed-point representation of fractional datatypes. We assume a processor having registers of b bits (including sign) and accumulators of at least $B = 2b + \lceil \log_2 R \rceil + 1$ bits, where R is the FFT radix and $\lceil a \rceil$ is the smallest integer that is larger than a. The numbers represented by the registers are in 2's complement representation and in the range $-1 \le x \le 1 - 2^{-(b-1)}$. The numbers represented by the accumulators are in the rage $-2^{\lceil \log_2 R \rceil + 1} \le x < 2^{\lceil \log_2 R \rceil + 1}$. The width of the data stored to memory is always of b bits.

Our focus is of fixed-radix, Cooley-Tukey, DIT-FFTs of radix-2 and radix-4. A generic model of a finite-word-length radix-2/radix-4 butterfly of the DIT-FFT is given in Figure 1.

In the DIT topology the inputs loaded from the memory are first multiplied by the Twiddle Factors (TFs), w_N^{kn}, then multiplied by the butterfly's coefficients $\gamma_{r,t}; r, t \in \{0, 1, \dots, R-1\}$, and then summed up within the butterfly before being stored back to the memory. The processing model that we will deal here with is a model that is most common to DSPs and dedicated FFT processors. In this model the inputs x_n and the TFs w_N^{kn} are represented by b bits per component (b bits for the real component and b bits for the imaginary component) and are within the range of $\left[-1\,, 1 - 2^{-(b-1)}\right]$. When multiplied, the multiplication is spanned over $2b + 1$ bits (recalling that the TF multiplication is a complex multiplication). Since in radix-2 and radix-4 FFTs the butterfly's internal coefficients, $\gamma_{r,t}$, belong to the sets $\{1, -1\}$ and $\{1, -1, j, -j\}\ ;\ j = \sqrt{-1}$ respectively, there are no truly multiplications within the butterfly. The bit-width of the butterfly's output can grow to span over up to B bits and then potentially scaled down by a factor of α, where we restrict α to be a power of 2. The scaled down butterfly output is quantized to b bits per component via rounding before being stored to memory.

The quantization model that we use here is the so-called Rounding-Half-Up (RHU) [15], which is also known as hardware-friendly-rounding and is being used in most digital signal processors and hardware implementations of digital signal processing functions. The mathematical function of RHU rounding to b bits is

$$y = Q[s] \triangleq 2^{-b} \cdot \lfloor s \cdot 2^b + 0.5 \rfloor \tag{1}$$

where $\lfloor a \rfloor$ is maximal integer lower than a and $s \in \left[-1{,}1 - 2^{-(b-1)}\right]$. The quantization error is $v = s - y$ and in the general case is modeled as an additive noise having uniform distribution [16]

$$v \sim U[-2^{-b}, 2^{-b}) \tag{2}$$

and is independent of s. As we deal here with finite-word-length, in fact v has a discrete distribution. However, for large enough b it is common to treat it as a zero mean continuous uniform distribution. As such its variance is

$$\sigma_v^2 = \frac{2^{-2(b-1)}}{12}. \tag{3}$$

In addition, throughout the FFT there are plenty of cases where all the TFs preceding a given butterfly are among the set

$$\mathcal{T}_1 \triangleq \{1, -1, j, -j\}\ ;\ j = \sqrt{-1}\,. \tag{4}$$

In such cases, the multiplication of a b-bits value $x \in \left[-1\,, 1 - 2^{-(b-1)}\right]$ by the TF $w \in \mathcal{T}_1$ would result in a $2b$-bits number, $t = w \cdot x$, that it's lower b bits are equal to zero. If all the TFs preceding a given butterfly are among the set $\mathcal{T}_1$, then the lower b bits of the butterfly's outputs, before down scaling, are also equal to zero. When such a number is scaled down by very few bits, the quantization noise does not obey to the uniform distribution anymore [16]. In this case we get a Random Variable (RV) having discrete distribution and non-zero mean. For example, in the case that such a number is shifted one bit to the right, the quantization noise ε_1 is distributed as

$$\varepsilon_1 = \begin{cases} 0 & w.p.\ 0.5 \\ -\frac{1}{2}2^{-(b-1)} & w.p.\ 0.5\,, \end{cases} \tag{5}$$

where the subscript 1 in ε_1 refers to the case of quantization noise generated by right shift of the b-bits number by one bit. The expected value of this noise equals $-2^{-(b-1)}/4$ and hence when dealing with Signal-to-Quantization-Noise-Ratios of those RVs we will relate to the noise power rather than to its variance. To distinguish the power from the variance we use the symbol ρ^2 for power. The expected value of the power of ε_1 RV then is

$$\rho_{\varepsilon_1}^2 = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \left(\frac{1}{2}2^{-(b-1)}\right)^2 = \frac{2^{-2(b-1)}}{8}. \tag{6}$$

As expected, this is larger than the variance of the zero mean uniformly distributed quantization noise of (3). In a similar way we can calculate the noise power of quantization noises that are generated due to the rounding after right shift of a b-bits number by q bits. In most FFT topologies and radices up to Radix-5, the right shifts are in the range of 0 to 3. Moreover, for right shifts of 4 and above the quantization noise power is very close to the variance of the zero mean uniform quantization noise of (3). Therefore, for our analytical derivations we use

$$\rho_{\varepsilon_q}^2 = \begin{cases} 0 & ; \quad q = 0 \\ \frac{1}{8}2^{-2(b-1)} & ; \quad q = 1 \\ \frac{3}{32}2^{-2(b-1)} & ; \quad q = 2 \\ \frac{11}{128}2^{-2(b-1)} & ; \quad q = 3 \\ \frac{1}{12}2^{-2(b-1)} & ; \quad q \geq 4\,. \end{cases} \tag{7}$$

In the sequel, we designate the set of butterflies that all their inputs were multiplied by TFs belonging to $\mathcal{T}_1$, as the $\mathcal{B}_1$ set or $\mathcal{B}_1$ butterflies.

III. SQNR OF A GENERIC BFP-FFT

By "generic BFP-FFT" we refer to a BFP-FFT that incorporates policy for down-scaling by right shifts at the outputs of the FFT stages, where the decision at which stages to scale down and by what factor are the policy parameters. In the following paragraphs we will relate to specific BFP scaling policies and will analyze their SQNR performance. We assume zero mean i.i.d. input sequence, $x(n)$, and that the quantization is regarded as an i.i.d. noise source. Moreover, multiple quantization noises at the input to a given butterfly that have been generated at earlier stages are mutually uncorrelated [9]. In order to derive the analytical expression of the SQNR, we will adopt the analysis strategy of Weinstein [9]. Let us relate to an input sequence of length N, $x(n)$, and a fixed-radix FFT of radix R. Define $M = \log_R N$, and α_m as the scale value at the output of the m^{th} stage, $m \in \{1, 2, \dots, M\}$, where we restrict α_m to be of the form $\alpha_m = 2^{-q_m}$ and q_m is a positive integer. We denote $x_m(n)$ as the array values at the output of the m^{th} stage, where $x_M(k) \triangleq X(k)$ is the FFT output, and $x_0(n) \triangleq x(n)$ is the FFT input. For a zero mean, i.i.d. sequence $x(n)$, the variance of the signal at the FFT output is given by

$$\sigma_{x_M}^2 = N\sigma_{x_0}^2 \prod_{m=1}^{M} \alpha_m^2 = N\sigma_{x_0}^2 2^{-2\sum_{m=1}^{M} q_m}\,. \tag{8}$$

The noise at the output of a given butterfly is composed of two components: the noise that is generated by that particular butterfly, which we call butterfly self-noise, and the noise that is propagated through the butterfly (noise that was generated at earlier stages), which we call propagated-noise. The propagated-noise power is multiplied by a factor of $R\alpha^2$ as each butterfly output is composed of the sum of R i.i.d. noise values and is multiplied by a scaling factor α. The self-noise, v, is the noise generated by the quantization at the butterfly output after being multiplied by α as depicted in Figure 1. Its variance is denoted as σ_v^2 (or power of ρ_v^2). Looking at the output noise of an M stages FFT, it is observed that the noise from the first stage propagates through the following M-1 stages, which results in accumulation of R^{M-1} such i.i.d. noise sources, each attenuated by a factor of $\prod_{m=2}^{M} \alpha_m^2$. The propagation of the noise from the second stage results in accumulation R^{M-2} such i.i.d. noise sources, each attenuated by a factor of $\prod_{m=3}^{M} \alpha_m^2$, and so on. The total output noise variance, σ_E^2, for an M stages FFT, assuming all the quantization operations are modeled as uniform RVs, $U[-2^{-b}, 2^{-b})$, is given by the following expression

$$\begin{aligned} \sigma_E^2 &= \sigma_v^2 \left(1 + \sum_{m=1}^{M-1} \prod_{i=m+1}^{M} R\alpha_i^2\right) \\ &= \sigma_v^2 \left(1 + \sum_{m=1}^{M-1} R^{M-m} \prod_{i=m+1}^{M} \alpha_i^2\right). \end{aligned} \tag{9}$$

In (9) it was assumed that the self-noise is a continuous RV and have the same PDF at all the butterflies. For b sufficiently large (e.g., $b = 16$) this assumption is commonly accepted. However, this is not the case for butterflies belonging to the $\mathcal{B}_1$ set in which their outputs are discrete RVs with Probability-Mass-Function (PMF) that depend on the number of right shifts took place at the butterfly output. The power of those noise sources is larger than that of the uniform RV, and hence they have negative effect on the quantization noise power at the FFT output. In order to be able to evaluate the effect of those noise sources, we want to incorporate their statistical model in the derivation of ρ_E^2.

Let us denote by β_m the fraction of the butterflies belonging to the $\mathcal{B}_1$ set at stage m, and by $\rho_{q_m}^2$ the self-noise power at the output of those butterflies. Using those notations, and relating to power-of-two FFTs, we can now re-write (9) as

$$\begin{aligned} \rho_E^2 = \sigma_v^2 &\sum_{m=1}^{M} R^{M-m+1} \prod_{i=m+1}^{M+1} \alpha_i^2 \\ &+ \sum_{m=1}^{M} \beta_m \left(\rho_{q_m}^2 - \sigma_v^2\right) R^{M-m+1} \prod_{i=m+1}^{M+1} \alpha_i^2\,, \end{aligned} \tag{10}$$

where we defined a virtual α_{M+1} set to $\alpha_{M+1} = 1/\sqrt{R}$.
The second term in (10) is a positive quantity that represents the increased output noise power caused by butterflies of the set $\mathcal{B}_1$. As we are dealing with power-of-two DIT FFTs, we can write the precise expression of β_m as a function of the radix R. This is easily extracted from the flow graphs of those FFTs and is equal to

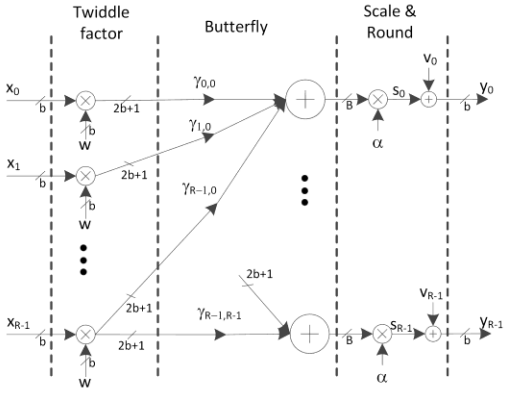


Figure 1. Generic model of DIT FFT Butterfly

$$\beta_m(R) = \begin{cases} R^{-(m-1)} & ;\ R > 2 \\ 1 & ;\ R = 2, m = 1 \\ R^{-(m-2)} & ;\ R = 2, m > 1\,. \end{cases} \quad (11)$$

Now we can plug β_m into (10) and get for $R = 2$

$$\begin{aligned} \rho_E^2 = \sigma_v^2 \sum_{m=1}^{M} R^{M-m+1} \prod_{i=m+1}^{M+1} \alpha_i^2 \\ + \left(\rho_{q_1}^2 - \sigma_v^2\right) R^M \prod_{i=2}^{M+1} \alpha_i^2 \\ + \sum_{m=2}^{M} \left(\rho_{q_m}^2 - \sigma_v^2\right) R^{M-2m+3} \prod_{i=m+1}^{M+1} \alpha_i^2 \end{aligned} \quad (12)$$

and for $R > 2$

$$\begin{aligned} \rho_E^2 = \sigma_v^2 \sum_{m=1}^{M} R^{M-m+1} \prod_{i=m+1}^{M+1} \alpha_i^2 \\ + \sum_{m=1}^{M} \left(\rho_{q_m}^2 - \sigma_v^2\right) R^{M-2m+2} \prod_{i=m+1}^{M+1} \alpha_i^2\,. \end{aligned} \quad (13)$$

Using (8), (12) and (13), the SQNR for a given scale pattern, $\boldsymbol{q} = [q_1, q_2, \dots, q_M]$, can be calculated by $\sigma_{x_M}^2 / \rho_E^2$ where assigning $\alpha_i = 2^{-q_i}$.

IV. SCALING POLICIES

In most FFT realizations, we wish to select a scaling policy that maximizes the SQNR under the constraint of zero-overflows. At the ideal BFP-FFT, the scaling policy is such that throughout the butterflies' computation, every butterfly's output is tested for an overflow before it is quantized down to b bits. If the real or the imaginary components of the butterfly output overflows, the entire stage is re-calculated where the butterflies' outputs are scaled down by q bits before being rounded to b bits and stored to memory. The value q is selected to guarantee that the scaled result does not overflow anymore. For example, if one of the absolute values of the real or imaginary butterfly's outputs is within the range $\left[1, 2 - 2^{-(b-1)}\right]$, the entire stage will be re-calculated while the butterflies' outputs will be scaled by one bit to the right ($q = 1$). If one of the absolute values of the real or imaginary butterfly's outputs is within the range $\left[2, 4 - 2^{-(b-1)}\right]$, the entire stage will be re-calculated while the butterflies' outputs will be scaled by two bits to the right and so on. The more common, fixed latency policy proposed by Shively [11] guarantees deterministic latency at the expense of decreased SQNR. In this policy, the decision by what factor to down-scale the outputs of stage m is taken based on the values of the outputs of stage $m-1$, which are guaranteed to fit in the range $\left[-1\,, 1 - 2^{-(b-1)}\right]$. While writing the outputs of stage $m-1$ to the memory, the processor finds the maximal absolute value among the real and imaginary components of the whole stage, which serves for the down-scaling decision for the next stage. The down-scaling criteria is similar to that being used at the ideal BFP-FFT, i.e., to guarantee zero overflow at the output of the next stage. Here, there is a need to consider the fact that the maximal absolute value at the next stage (stage m) butterflies' output would grow by a factor that is between 1 and $\sqrt{2}R$ relative the outputs of the current stage (stage $m-1$). In order to formalize this, let us define $x_m^c(n)$ for $n \in \{0, 1, \dots, N-1\}$ as

$$\begin{aligned} x_m^c(2n) &= real\left(x_m(n)\right) \\ x_m^c(2n+1) &= imag\left(x_m(n)\right) \end{aligned} \quad (14)$$

and

$$\tilde{x}_m = \max_n \{|x_m^c(n)|\}\,. \quad (15)$$

Using those, the scaling policy of the practical BFP-FFT can be written as

$$q_m = \begin{cases} 0 & ;\ \tilde{x}_{m-1} < \frac{1}{\sqrt{2}R} \\ 1 & ;\ \frac{1}{\sqrt{2}R} \le \tilde{x}_{m-1} < \frac{2}{\sqrt{2}R} \\ 2 & ;\ \frac{2}{\sqrt{2}R} \le \tilde{x}_{m-1} < \frac{4}{\sqrt{2}R} \\ \vdots & \\ \lceil log_2(R) \rceil + 1 & ;\ \frac{1}{\sqrt{2}} \le \tilde{x}_{m-1} \end{cases}. \quad (16)$$

V. SQNR CALCULATION

It is now clear that the SQNR at the FFT output of a particular realization of the FFT depends on the scale pattern that has been used throughout this realization. Each scale pattern, $\boldsymbol{q}$, is associated with a resultant SQNR. We adopt Weinstein's definition for "theoretical" SQNR as the weighted sum of the SQNR per scale pattern over all possible patterns [9]. The probability of a scale pattern depends solely on the PDF of the input sequence and the scaling policy. In the sequel we will derive the scale patterns probabilities as well as the SQNR of the practical BFP-FFT and of the ideal BFP-FFT algorithms for Gaussian input sequences.

A. *Scale patterns probabilities of practical BFP-FFT*

We start with the derivation of the probabilities of scale patterns. Given the practical BFP-FFT's scaling policy, the probability that there will be exactly $q > 0$ right shifts at stage m is equal to

$$\begin{aligned} Pr(q_m = q) &= Pr\left(\frac{2^{q-1}}{\sqrt{2}R} \le \tilde{x}_{m-1} \le \frac{2^q}{\sqrt{2}R}\right) \\ &= Pr\left(-\frac{2^q}{\sqrt{2}R} \le \underset{n}{all}\{x_{m-1}^c(n)\} \le \frac{2^q}{\sqrt{2}R}\right) \\ &\quad - Pr\left(-\frac{2^{q-1}}{\sqrt{2}R} \le \underset{n}{all}\{x_{m-1}^c(n)\} \le \frac{2^{q-1}}{\sqrt{2}R}\right) \end{aligned} \quad (17)$$

whereas for $q = 0$

$$Pr(q_m = 0) = Pr\left(\tilde{x}_{m-1} \le \frac{1}{\sqrt{2}R}\right). \quad (18)$$

By the assumption that the input sequence, $x_{m-1}^c(n); n \in$

$\{0, 1, \dots, 2N-1\}$ is an i.i.d. sequence, (17) and (18), can be written as

$$Pr(q_m = q) = \left[Pr\left(-\frac{2^q}{\sqrt{2}R} \le x_{m-1}^c(n) \le \frac{2^q}{\sqrt{2}R}\right)\right]^{2N} - \left[Pr\left(-\frac{2^{q-1}}{\sqrt{2}R} \le x_{m-1}^c(n) \le \frac{2^{q-1}}{\sqrt{2}R}\right)\right]^{2N} \quad (19)$$

whereas for $q = 0$

$$r(q_m = 0) = \left[Pr\left(-\frac{1}{\sqrt{2}R} \le x_{m-1}^c(n) \le \frac{1}{\sqrt{2}R}\right)\right]^{2N}. \quad (20)$$

We now define the following auxiliary variables

$$Q_m = \sum_{i=1}^{m} q_i \ ; \ m \in \{1, 2, \dots, M\} \ , \ \ Q_0 = 1 \quad (21)$$

and

$$T_m = 2^{-2Q_m}. \quad (22)$$

Using those, the variance of the sequence at the output of the m^{th} stage is

$$\sigma_{x_m}^2 = \sigma_{x_0}^2 R^m T_m \quad (23)$$

and the variance of the real and imaginary individual components at the output of the m^{th} stage is $\sigma_{x_0}^2 R^m T_m/2$. For an i.i.d complex Gaussian input sequence, $x_0^c(n) \sim N\left(0, \sigma_{x_0}^2/2\right)$; $n \in \{0, 1, \dots, 2N-1\}$, it can be shown that all the intermediate sequences $x_m^c(n)$, $m \in \{1, 2, \dots, M\}$ are also Gaussian i.i.d [9]. Therefore, the probability that the outputs of the m^{th} stage would be shifted by exactly $q_m > 0$ right shifts, given that there were accumulated Q_{m-1} right shifts at the stages preceding stage m is

$$Pr\left(q_m \mid Q_{m-1}; \sigma_{x_0}^2\right) = \left[erf\left(\frac{2^{q_m}}{\sigma_{x_0}\sqrt{2R^{m+1}T_{m-1}}}\right)\right]^{2N} - \left[erf\left(\frac{2^{q_m-1}}{\sigma_{x_0}\sqrt{2R^{m+1}T_{m-1}}}\right)\right]^{2N} \quad (24)$$

and the probability that there would be no right shifts ($q_m = 0$) is given by

$$r\left(q_m = 0 \mid Q_{m-1}; \sigma_{x_0}^2\right) = \left[erf\left(\frac{1}{\sigma_{x_0}\sqrt{2R^{m+1}T_{m-1}}}\right)\right]^{2N} \quad (25)$$

where $erf(x)$ is defined by

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \quad (26)$$

B. *Scale patterns probabilities of ideal BFP-FFT*

At the scaling policy of the ideal BFP-FFT there are no per-stage scaling pre-decisions. An FFT stage is calculated without scaling and throughout the calculations, if any of the stage's outputs overflows, the whole stage is re-calculated while the outputs are down-scaled before being written to memory. Note that in the ideal policy there may be multiple re-calculation of the same stage if the strategy is to initiate the re-calculation upon the first detected overflowed value. Different strategies may eliminate the multi re-calculations of the same stage, for example set the scale value to the maximal scale upon the detection of the first overflow, or always calculate the stage to its end and if overflows have been detected, set the scale value according the largest magnitude among the overflowed values. Some strategies suffer degradations in the SQNR performance due to potential mismatch between the scale value and the actual overflowed value. Nevertheless, here, for the sake of SQNR comparison, we assume a strategy that determine scale value according to the largest magnitude output sample, and hence no performance loss is involved.

As opposed to the practical case, at which the scale decision for stage m depends on the outputs of stage $m-1$ after being scaled down, the scale decision of the ideal BFP-FFT depend on the outputs of stage m before being scaled down. Let us denote those values as $s_m(n)$, i.e.

$$x_m(n) = \alpha_m s_m(n) \quad (27)$$

and define $s_m^c(n)$ and $\tilde{s}_m$ in analogous to (14) and (15) as

$$s_m^c(2n) = real\left(s_m(n)\right) \\ s_m^c(2n+1) = imag\left(s_m(n)\right) \quad (28)$$

and

$$\tilde{s}_m = \max_n\{|s_m^c(n)|\}. \quad (29)$$

Now, the SQNR analysis using the ideal BFP-FFT policy follows the steps of the analysis of the practical BFP-FFT scheme. The output signal variance and the output noise power follow (8) and (10) respectively. The probability that there will be exactly $q > 0$ right shifts at stage m is equal to

$$Pr(q_m = q) = Pr(2^{q-1} \le \tilde{s}_m \le 2^q) = Pr\left(-2^q \le \underset{n}{all}\{s_m^c(n)\} \le 2^q\right) - Pr\left(-2^{q-1} \le \underset{n}{all}\{s_m^c(n)\} \le 2^{q-1}\right), \quad (30)$$

and the probability that there will be no right shifts at stage m, i.e. $q = 0$, is

$$Pr(q_m = 0) = Pr(\tilde{s}_m \le 1) = Pr\left(-1 \le \underset{n}{all}\{s_m^c(n)\} \le 1\right). \quad (31)$$

Under i.i.d. Gaussian input assumption, we get for $q_m > 0$

$$Pr\left(q_m \mid Q_{m-1}\,;\,\sigma_{x_0}^2\right) = \left[erf\left(\frac{2^{q_m}}{\sigma_{x_0}\sqrt{R^m T_{m-1}}}\right)\right]^{2N} - \left[erf\left(\frac{2^{q_m-1}}{\sigma_{x_0}\sqrt{R^m T_{m-1}}}\right)\right]^{2N}, \quad (32)$$

and for $q_m = 0$

$$Pr\left(q_m = 0 \mid Q_{m-1}\,;\,\sigma_{x_0}^2\right) = \left[erf\left(\frac{1}{\sigma_{x_0}\sqrt{R^m T_{m-1}}}\right)\right]^{2N}. \quad (33)$$

C. SQNR calculation

We use the per-stage probabilities to calculate the probability of a specific scale pattern, $\boldsymbol{q} = [q_1, q_2, \ldots, q_M]$,

$$Pr\left(\boldsymbol{q}; \sigma_{x_0}^2\right) = Pr\left(q_1; \sigma_{x_0}^2\right) \prod_{m=2}^{M} Pr\left(q_m | Q_{m-1}; \sigma_{x_0}^2\right) \quad (34)$$

and the output SQNR is calculated by the weighted sum of the SQNRs per scale pattern as

$$SQNR = \sum_{\boldsymbol{q}} Pr\left(\boldsymbol{q}; \sigma_{x_0}^2\right) \cdot SQNR\left(\boldsymbol{q}, \sigma_{x_0}^2\right) = \sum_{\boldsymbol{q}} Pr\left(\boldsymbol{q}; \sigma_{x_0}^2\right) \cdot \frac{\sigma_{x_M}^2\left(\boldsymbol{q}, \sigma_{x_0}^2\right)}{\rho_E^2(\boldsymbol{q})}. \quad (35)$$

In (35) the expression $Pr\left(\boldsymbol{q}; \sigma_{x_0}^2\right)$ is calculated by (34), $\sigma_{x_M}^2\left(\boldsymbol{q}, \sigma_{x_0}^2\right)$ is calculated by (8) and $\rho_E^2(\boldsymbol{q})$, with $\alpha_i = 2^{-q_i}$, is calculated by (12) or (13) for Radix-2 and Radix-4 respectively.

VI. RESULTS

The derived models of the SQNR for the practical and the ideal BFP-FFT have been validated against simulation. The model and the simulation results for 16-bit datatype ($b = 16$) and Gaussian i.i.d input with standard deviation of $\sigma_{x_0} = 0.15$ are shown in Figure 2 and Figure 3 for radix-2 and radix-4 respectively. For the simulation results we have averaged the SQNR of 1000 FFT runs per FFT length. As can be seen, there is a very good match between the simulation results and the derived model. The gap between the refined statistical model (that incorporate the refinement for $\mathcal{B}_1$ butterflies) and the simulation result for the practical BFP-FFT is in the order of 0.2dB. The results for the ideal BFP-FFT are not shown in the figures since the model has almost perfect match to the simulation result with gaps that are in the order of 0.05dB.

In Figure 2 and Figure 3 we can also see the effect of the refined statistical model for the $\mathcal{B}_1$ butterflies. The model neglecting the effects of the $\mathcal{B}_1$ butterflies, for radix-2 BFP-FFT, is optimistic by about 0.5dB for the practical BFP-FFT and by about 1dB for radix-4.

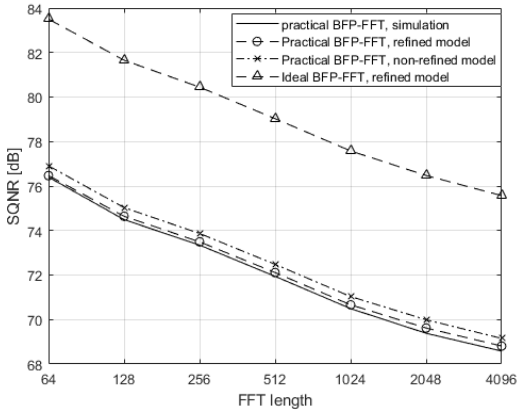


Figure 2. Radix-2 Practical BFP-FFT

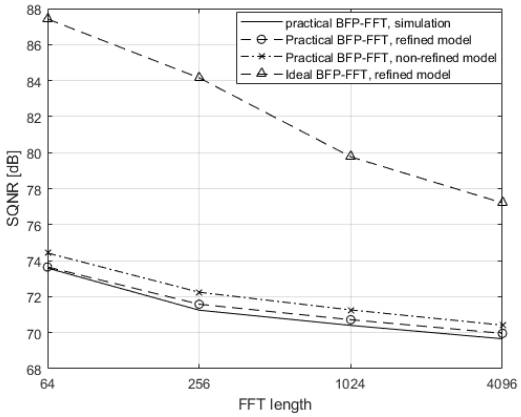


Figure 3. Radix-4 Practical BFP-FFT

One of the main goals of the paper is to provide an analytical tool that enables the prediction of the SQNR penalty one needs to pay for getting fixed latency BFP-FFT. This penalty is clearly seen for radix-2 and radix-4 in Figure 2 and Figure 3 respectively. We see that such a penalty is in the order of 6dB when the number of stages is above five, and grows up to 13.5dB for lower number of stages as seen at the case of 64 points radix-4 FFT.

Another interesting observation that the model reveals relates to the comparison of the SQNR between radix-2 and radix-4 BFP-FFT implementations. It is well known that from complexity perspective, the radix-4 has advantages over radix-2 (at least in the number of multiplications). From the results in Figure 2 and Figure 3, we can also see that radix-4 have better SQNR in the ideal BFP-FFT implementation. We get 4dB advantage for 64-points FFT down to about 2dB advantage for 4096-points FFT. However, for the practical BFP-FFT we see an opposite behavior. The radix-2 practical BFP-FFT results in 2.8dB better SQNR for 64-point FFT, down to 1.2dB better SQNR for 4096-points FFT.

VII. CONCLUSIONS

In this paper, we refined the analytical model of the finite-word-length-effects of Cooley Tukey DIT BFP-FFT to incorporate butterflies belonging to the $\mathcal{B}_1$ set, as well as extended the model for the commonly used practical BFP-FFT. The refined analytical model was validated against simulation and found highly accurate for ideal and practical BFP-FFTs. The model enables to accurately predict the SQNR for the practical BFP-FFT and the performance degradation compared to the ideal BFP-FFT scheme.

The analysis covers DIT-FFT for radix-2 and radix-4, but can be easily adapted to DIF FFT topologies and be extended for non-power-of-2 BFP-FFTs as well as for mixed radices, such as the ones used in LTE wireless modems.

REFERENCES

[1] J. M. Cioffi, at al., "Very-high-speed digital subscriber lines," *IEEE Communications Magazine*, vol. 37, no. 4, pp. 72-79, 1999.

[2] B. F. Frederiksen and R. Prasad, "An overview of OFDM and Related Techniques Towards Development of Future Wireless Multimedia Communications," in *IEEE Proc. Radio and Wireless Conference*, Boston, pp. 19-22, 2002.

[3] N. Cvijetic, "OFDM for Next-Generation Optical Access Networks," *IEEE Journal of Lightwave Technology*, vol. 30, no. 4, pp. 384-398, 2012.

[4] A. V. Oppenheim and C. J. Weinstein, "Effects of Finite Register Length in Digital Filtering and the Fast Fourier Transform," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 957-976, 1972.

[5] W.-H. Chang and N. Q. Truong, "On the Fixed-Point Accuracy Analysis of FFT Algorithms," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4973-4682, 2008.

[6] P. Gupta, "Accurate Performance Analysis of a Fixed Point FFT," in *Twenty Second National Conference on Communication (NCC)*, Guwahati, 2016.

[7] A. Monther and K. Zsolk, "Analysis of Quantization Noise in FFT Algorithms for Real-Valued Input Signals," in *International Conference on Radioelektronika*, Kosice, 2022.

[8] *LTE-A; Evolved Universal Terrestrial Radio Access (E-UTRA), Physical Channels and Modulation*, 3GPP TS 36.211, 2011.

[9] C. J. Weinstein, "Quantization Effects in Digital Filters," M.I.T. Lincoln Lab. Tech. Rep. 468, ASTIA doc. DDC AD-706862, 1969.

[10] Tran-Thong and B. Liu, "Fixed-Point Fast Fourier Transform Error Analysis," *IEEE Transactions on Acoustic, Specch, and Signal Processing*, vol. 24, no. 6, pp. 563-573, 1976.

[11] R. R. Shively, "A Digital Processor to Generate Spectra in Real Time," *IEEE Transactions on Computers*, Vols. C-17, no. 5, pp. 485-491, 1968.

[12] H. G. Kim, K. T. Yoon, J. S. Youn, and J. R. Choi, "8K-point Pipelined FFT/IFFT with Compact Memory for DVB-T using Block Floating-point Scaling Technique," in *International Symposium on Wireless Pervasive Computing (ISWPC)*, Melbourne, pp. 41-47, 2009.

[13] S. J. Huang and S. G. Chen, "A High-Parallelism Memory-Based FFT Processor with high SQNR and novel addressing scheme," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, Montreal, pp. 2671-2674, 2016.

[14] P. D. Welch, "A Fixed-Point Fast Fourier Transform Error Analysis," *IEEE Transactions on audio and Electroacoustics*, vol. 17, no. 2, pp. 151-157, 1969.

[15] L. Xia, M. Athonissen, M. Hochstenbach, and B. Koren, "Improved Stochastic Rounding," *arXiv , 2020, Available: https://arxiv.org/abs/2006.00489.*

[16] B. Widrow, I. Kollar, and M.-C. Liu, "Statistical theory of Quantization," *IEEE Transactions on Instrumentation and Measurement*, vol. 45, no. 2, pp. 353-361, 1996.

Minecraft of System Modeling

Pavel Loskot
ZJU-UIUC Institute
Haining, China
pavelloskot@intl.zju.edu.cn

***Abstract*—The best selling computer game of all times, Minecraft, represents the world as discrete blocks. The Minecraft-like worlds may be unknowingly created by many mathematical models of the real-world systems, when their inputs and outputs are discretized. This paper investigates system modeling and identification with noisy, discretized, but otherwise static inputs and outputs. Such a scenario occurs, for example, when configuring and measuring the system is time-consuming and costly. The task is to infer the model parameters from a limited number of input-output measurements. It is shown that, in this setting, the traditional least-squares model fitting is ineffective. A better strategy is to first accurately estimate the static input and output values, and then obtain the model parameters by inverting the model numerically by solving an underlying set of equations for the same number of unknown model parameters. These results have direct implications on creating and interpreting mathematical models of systems, and even physical laws, when the noisy measurements are implicitly or explicitly discretized.**

***Keywords*—Linear model; Mean-square error; Minecraft; Quantization; Parameter estimation; System identification.**

I. Introduction

Mathematical models are used extensively in many applications. The models are usually represented by the sets of parameterized equations describing the model input-output relationships. The aim of model identification is to recover the model parameters from the noisy measurements of its inputs and outputs. These measurements may be explicitly or implicitly quantized. The former is used to reduce the storage and transmission requirements, and to speed-up computations at the expense of loosing some information and accuracy. The implicit quantization is more subtle, and it occurs when the resolution of measured samples is insufficient, for example, due to the use of inexpensive measuring equipment.

Simply inverting the model in order to recover the model parameters from the measurements of its inputs and outputs is often unacceptable. The model inversion tends to greatly amplify the measurement noises, which leads to large estimation errors [1]. The model-based parameter estimation methods are often used to obtain the optimum and numerically efficient estimators in the presence of strong measurement noises. However, for model identification [2] and supervised machine learning [3], an alternative strategy can be adopted. In particular, the input and output values can be estimated independently from their noisy, and possibly discretized measurements. For static values, this corresponds to estimating unknown constants in additive noises. If the estimators used are unbiased and consistent, the measurement noises can be sufficiently suppressed, so the model inversion is acceptable to accurately infer the model parameters.

The paper [4] is one of the earliest studies on estimating the state of dynamic linear systems from quantized measurements. The authors demonstrated that Kalman filtering is still effective even under these conditions. This problem was considered again in [5] as a joint design of the quantizer and the estimator. The classical textbook [2] covers a wide range of topics in adaptive filtering including system identification and adaptive filter design with quantized inputs. The paper [6] investigates the optimum techniques for signal detection and estimation, and evaluates the corresponding performance losses due to uniform signal quantization. The confidence intervals of the discretized likelihood-based estimators with quantized inputs were studied in [7]. The encoding and decoding schemes for quantized random processes were designed in [8] to enable their efficient transmissions under the age-of-information constraints. The Cramér-Rao bounds for estimating the parameters from quantized measurements were derived in [9].

In this paper, we consider the problem of identifying the model parameters from quantized noisy measurements of both the model inputs and outputs. The model inputs and outputs are assumed to be static, so their values can be inferred with a high accuracy from a sufficient number of measurements assuming the consistent and unbiased estimators. The model parameters are then obtained by solving a set of linear or non-linear equations. It is also shown that the traditional least squares fitting of the model to the input and output data is much less effective, when the input and output measurements are noisy and quantized. This is also an important issue, for example, in supervised machine learning.

The following notations are adopted in the paper: $\mathrm{Av}[\cdot] = (1/T)\int_{-T/2}^{T/2}(\cdot)\,\mathrm{d}t$, and, $\mathrm{Av}[\cdot] = (N+1)^{-1}\sum_{i=-N/2}^{N/2}(\cdot)$, are the time-averaging (arithmetic average) operators in continuous and discrete time, respectively, $\mathrm{E}[\cdot]$ is the statistical expectation, $\boldsymbol{x}$ denotes a column vector, whereas $\boldsymbol{X}$ denotes a matrix, $(\cdot)^T$ and $(\cdot)^{-1}$ denote the matrix transpose and inverse, respectively, $\langle\cdot,\cdot\rangle$ denotes the dot-product, $\dot{f}$ is the first derivative of function, f, $\lfloor\cdot\rfloor$, $\lceil\cdot\rceil$, and $\mathrm{sign}(\cdot)$ are the floor function, the ceiling function, and the sign function, respectively, and $\mathbb{R}$ and $\mathbb{Z}$ represent the sets of real numbers and integers, respectively.

The rest of the paper is organized as follows. Section II outlines system model with uniformly, and also binary quantized inputs and outputs. The estimation of model parameters is described in Section III. The estimator variances are studied in Section IV. Discussion and future work are in Section V.

II. System Model

A general parameterized model with multiple inputs and outputs (MIMO) is shown in Figure 1. Such a model can be succinctly described by a single equation,

$$f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a}) = 0 \tag{1}$$

relating the model inputs, $\boldsymbol{x}$, outputs, $\boldsymbol{y}$, and a given set of model parameters, $\boldsymbol{a}$. Importantly, it is assumed that the input as well as output measurements of model (1) are first quantized and de-noised, before estimating the parameters, $\boldsymbol{a}$.

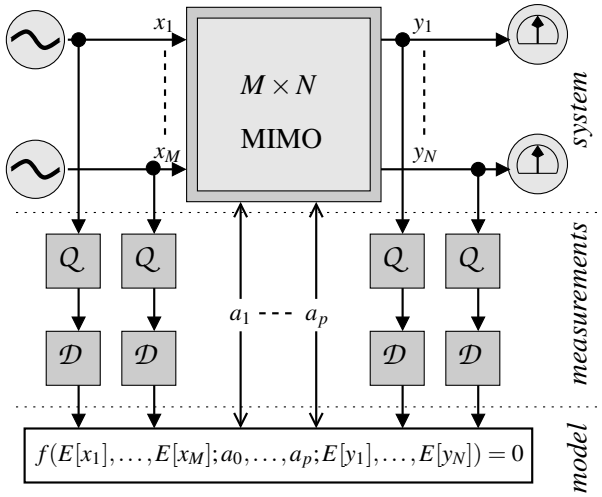


Figure 1. Modeling and measurements of a static $(M \times N)$ MIMO system.

Note that, here, the system modeling assumes the expected values of the inputs and outputs. In practice, measuring the statistical means can be problematic, when the random processes are non-stationary or non-ergodic [10]. The measuring instruments usually report the time-averaged values over a certain time-window. On the other hand, the expected values are more a theoretical concept, which is used, for example, when deriving the estimators of random signals to minimize the given risk. However, under the law of large numbers, the expectations can be replaced by the time averages. These differing views and assumptions can be reconciled by assuming the statistical and time averaging at the same time, i.e., by assuming, $\mathrm{Av}[\mathrm{E}[\cdot]] = \mathrm{E}[\mathrm{Av}[\cdot]]$. Depending on the type of a random process, $x(t)$, different averages are related as:

$$\begin{aligned}
\mathrm{E}[x] = \mathrm{Av}[\mathrm{E}[x]] = \mathrm{Av}[x] \quad &\Leftrightarrow \quad \text{ergodic \& stationary,} \\
\mathrm{E}[x] \neq \mathrm{Av}[\mathrm{E}[x]] = \mathrm{Av}[x] \quad &\Leftrightarrow \quad \text{ergodic \& non}-\text{stationary,} \\
\mathrm{E}[x] = \mathrm{Av}[\mathrm{E}[x]] \neq \mathrm{Av}[x] \quad &\Leftrightarrow \quad \text{non}-\text{ergodic \& stationary,} \\
\mathrm{E}[x] \neq \mathrm{Av}[\mathrm{E}[x]] \neq \mathrm{Av}[x] \quad &\Leftrightarrow \quad \text{non}-\text{ergodic \& non}-\text{stationary.}
\end{aligned} \tag{2}$$

A. Linear SISO model

For the sake of notational simplicity, consider a single-input, single-output (SISO) model.

The linear SISO model is described by a linear combination of p basis functions, $\phi_i(x)$, i.e.,

$$y = a_0 + \sum_{i=1}^{p} a_i \phi_i(x). \tag{3}$$

If the functions, $\phi_i(x)$, are mutually orthogonal, i.e., the dot-product, $\langle \phi_i, \phi_j \rangle \neq 0$, for $\forall i \neq j$, then p is also the dimension (rank) of the linear model. The n output measurements, y_i, collected at n input values, x_i, are related as,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \phi_1(x_1) & \cdots & \phi_p(x_1) \\ \vdots & \vdots & & \vdots \\ 1 & \phi_1(x_n) & \cdots & \phi_p(x_n) \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} \tag{4}$$

$$\boldsymbol{y} = \boldsymbol{\Phi}(\boldsymbol{x}) \cdot \boldsymbol{a}.$$

The basis functions are generally non-linear, however, they can be linearized about a chosen value, x_0, as,

$$\phi_i(x) \doteq \phi_i(x_0) + \dot{\phi}_i(x_o)(x - x_0). \tag{5}$$

Such linear approximations can be also defined in multiple dimensions [11]. The caveat is that the approximation (5) is only valid in the vicinity of x_0, and choosing the suitable value can be problematic. For example, if linear model (3) represents a polynomial regression, then it can be rewritten assuming the linearized basis functions as,

$$y = a_0 + \sum_{i=1}^{p} a_i (A_i x + B_i) \tag{6}$$

where $A_i = \dot{\phi}_i(x_0)$, and, $B_i = \phi_i(x_0) - \dot{\phi}_i(x_0)x_0$.

B. Quantized measurements

The measurements are quantized for various reasons. For instance, the explicitly quantized values require less memory for storage, and the numerical computations become faster to perform. The implicit quantization occurs when the resolution of the measurements is insufficient with respect to a given modeling application. The most common is a uniform quantization having the equidistant quantization intervals of length, Δ, i.e.,

$$\check{x} = Q(x) = \left\lfloor \frac{x - \Delta/2}{\Delta} \right\rfloor + 1 \in \mathbb{Z} \tag{7}$$

so that the quantization error, $\varepsilon_\Delta = x - \Delta\check{x}$, and,

$$\Delta(\check{x} - 1/2) \leq x < \Delta(\check{x} + 1/2). \tag{8}$$

Note also that, $\lfloor x \rfloor + 1 \neq \lceil x \rceil$, for the integer values of the argument. In addition, the quantized values are often bounded to a finite set of integers between, $-\check{x}_{\max}$, and, $\check{x}_{\max}$.

Alternatively, the binary quantization,

$$\check{x} = Q_2(x) = \mathrm{sign}(x) \in \{-1, +1\} \tag{9}$$

can be sufficient in some applications.

The issue with implicit quantization due to insufficient resolution is illustrated in Figure 2, assuming a linear system, $y = 3x/2$, and the uniform quantization with $\Delta = 1/2$. It

can be observed that, the model having only the quantized inputs, $y = aQ(x)$, is nearly identical to the unquantized model, $y = ax$. However, when both the input and the output are quantized, a formerly linear model becomes a staircase function (red dashed line), $Q(y) = aQ(x)$. In this case, only one noise-free measurement is necessary to determine the constant, a. If such a measurement is taken at points, A, B, or C, the proportionality constant is inferred to be equal to 1, $5/4$, or $7/4$, respectively. Consequently, the implicit or explicit quantization of the output values have a severe impact on identifying the model parameters.

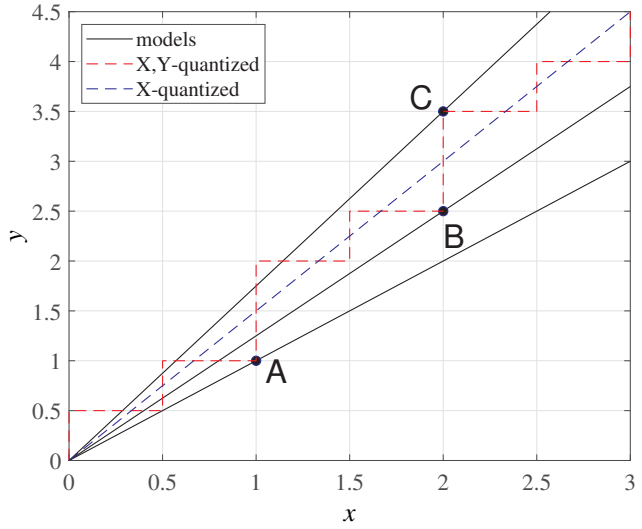


Figure 2. The consequences of the input-output uniform quantization on modeling linear SISO systems.

III. Estimating Model Parameters

Assume that n noisy measurements can be written as,

$$\begin{aligned} y_i &= \bar{y} + \varepsilon_{yi} \\ x_i &= \bar{x} + \varepsilon_{xi} \end{aligned} \tag{10}$$

where the additive noises, ε_{yi}, and, ε_{xi}, have zero means, i.e., $\mathrm{E}[y_i] = \bar{y}$, and, $\mathrm{E}[x_i] = \bar{x}$. Even when the measurements at different time instances can be assumed to be independent, the input-output correlations,

$$\mathrm{E}[x_i y_i] = \bar{x}\bar{y} + \mathrm{E}[\varepsilon_{xi}\varepsilon_{yi}] \tag{11}$$

are affected by the noise covariances, $\mathrm{E}[\varepsilon_{xi}\varepsilon_{yi}] \neq 0$.

Provided that the measurements are noisy, the input-output relationship (4) can only be satisfied approximately. The over-determined linear systems with $n \gg p$ measurements can be solved by considering the least-squares (LS) model fitting. The closed-form expression for the LS estimate of the model parameters is well-known, i.e., [12]

$$\hat{\boldsymbol{a}}_{\mathrm{LS}} = \left(\boldsymbol{\Phi}^T(\boldsymbol{x})\boldsymbol{\Phi}(\boldsymbol{x})\right)^{-1}\boldsymbol{\Phi}^T(\boldsymbol{x})\boldsymbol{y}. \tag{12}$$

Substituting the noisy measurements (10) into (12), while also assuming a linearization of the basis functions (5) about the mean, $\bar{x}$, the resulting linear model (4) can be written as,

$$\boldsymbol{y} = \left[\mathbf{1}_{(n,1)} \mid \bar{\boldsymbol{\Phi}}(\bar{x}) + \boldsymbol{\varepsilon}_x \cdot \dot{\boldsymbol{\phi}}^T(\bar{x})\right] \cdot \boldsymbol{a} \tag{13}$$

where $\mathbf{1}_{(n,1)}$ is the all-ones column vector, the constant matrix, $\bar{\boldsymbol{\Phi}}(\bar{x})$, has identical rows with the elements, $\phi_i(\bar{x})$, the column vector, $\boldsymbol{\varepsilon}_x$, contains additive noises, ε_{xi}, at the model input, and the constant column vector, $\dot{\boldsymbol{\phi}}(\bar{x})$, has the elements, $\dot{\phi}_i(\bar{x})$.

In order to obtain an insight into the LS solution of (13) for the model parameters, $\boldsymbol{a}$, consider the LS sum over the n measurements, i.e.,

$$\mathrm{LS}(a_0, \boldsymbol{a}) = \sum_{i=1}^{n}\left(y_i - a_0 - \left(\dot{\boldsymbol{\phi}}\varepsilon_{xi} + \boldsymbol{\phi}\right)^T \cdot \boldsymbol{a}\right)^2 \tag{14}$$

where the parameter, a_0, was taken out of the p-element vector, $\boldsymbol{a}$, $\boldsymbol{\phi}$ represents the row of the matrix, $\bar{\boldsymbol{\Phi}}(\bar{x})$, transposed to become a column vector, and let the vector of derivatives, $\dot{\boldsymbol{\phi}}(\bar{x}) \equiv \dot{\boldsymbol{\phi}}$. Note that both vectors, $\dot{\boldsymbol{\phi}}$, and, $\boldsymbol{\phi}$, are independent of the index, i. The model parameters minimizing the LS value are the solution of the set of linear equations, i.e.,

$$\begin{aligned} \frac{\partial}{\partial a_0}\mathrm{LS}(\hat{a}_0, \hat{\boldsymbol{a}}) &= 0 \\ \frac{\partial}{\partial \boldsymbol{a}}\mathrm{LS}(\hat{a}_0, \hat{\boldsymbol{a}}) &= \mathbf{0}. \end{aligned} \tag{15}$$

After some lengthy, but otherwise straightforward manipulations, we get,

$$\hat{a}_0 = \mathrm{Av}[y_i] - \left(\dot{\boldsymbol{\phi}}\mathrm{Av}[\varepsilon_{xi}] + \boldsymbol{\phi}\right)^T \hat{\boldsymbol{a}} \tag{16}$$

where $\mathrm{Av}[y_i] = (1/n)\sum_{i=1}^{n} y_i$, and, $\mathrm{Av}[\varepsilon_{xi}] = (1/n)\sum_{i=1}^{n}\varepsilon_{xi}$. Noticing that, $y_i - \mathrm{Av}[y_i] = \varepsilon_{yi}$, we obtain the solution for $\boldsymbol{a}$, which can be substituted into (16), i.e.,

$$\begin{aligned} &\dot{\boldsymbol{\phi}}\mathrm{Av}[\varepsilon_{xi}\varepsilon_{yi}] + \boldsymbol{\phi}\mathrm{Av}[\varepsilon_{yi}] = \\ &\left(\mathrm{Av}\left[\left(\dot{\boldsymbol{\phi}}\varepsilon_{xi} + \boldsymbol{\phi}\right)\left(\dot{\boldsymbol{\phi}}\varepsilon_{xi} + \boldsymbol{\phi}\right)^T\right] - \mathrm{Av}\left[\dot{\boldsymbol{\phi}}\varepsilon_{xi} + \boldsymbol{\phi}\right]\mathrm{Av}\left[\dot{\boldsymbol{\phi}}\varepsilon_{xi} + \boldsymbol{\phi}\right]^T\right)\boldsymbol{a}. \end{aligned} \tag{17}$$

The right-hand side of (17) can be further simplified as,

$$\dot{\boldsymbol{\phi}}\mathrm{Av}[\varepsilon_{xi}\varepsilon_{yi}] + \boldsymbol{\phi}\mathrm{Av}[\varepsilon_{yi}] = \dot{\boldsymbol{\phi}}\dot{\boldsymbol{\phi}}^T\mathrm{Av}\left[(\varepsilon_{xi} - \bar{\varepsilon}_x)^2\right]\boldsymbol{a} \tag{18}$$

where $\bar{\varepsilon}_x = \mathrm{Av}[\varepsilon_{xi}]$. Finally, the LS estimates of the model parameters are then computed as,

$$\hat{\boldsymbol{a}} = \left(\dot{\boldsymbol{\phi}}\dot{\boldsymbol{\phi}}^T\right)^{-1}\dot{\boldsymbol{\phi}}\frac{\mathrm{Av}[\varepsilon_{xi}\varepsilon_{yi}]}{\mathrm{Av}[(\varepsilon_{xi} - \bar{\varepsilon}_x)^2]} + \left(\dot{\boldsymbol{\phi}}\dot{\boldsymbol{\phi}}^T\right)^{-1}\boldsymbol{\phi}\frac{\mathrm{Av}[\varepsilon_{yi}]}{\mathrm{Av}[(\varepsilon_{xi} - \bar{\varepsilon}_x)^2]}. \tag{19}$$

For a large number of samples, $n \gg 1$, $\mathrm{Av}[\varepsilon_{yi}] \doteq 0$, and the final expression for estimating the model parameters becomes,

$$\hat{\boldsymbol{a}} = \left(\dot{\boldsymbol{\phi}}\dot{\boldsymbol{\phi}}^T\right)^{-1}\dot{\boldsymbol{\phi}}\frac{\mathrm{Av}[\varepsilon_{xi}\varepsilon_{yi}]}{\mathrm{Av}[(\varepsilon_{xi} - \bar{\varepsilon}_x)^2]}. \tag{20}$$

As an illustrative example, assume a simple linear SISO model, $y_i = a_1 x_i + a_0$, with $p = 2$ parameters. Assuming (20), the LS estimates of the model parameters are,

$$\begin{aligned} \hat{a}_0 &= \bar{y} - \hat{a}_1\bar{x} \\ \hat{a}_1 &= \frac{\mathrm{Av}[(y_i - \bar{y})(x_i - \bar{x})]}{\mathrm{Av}[(x_i - \bar{x})^2]} \end{aligned} \tag{21}$$

where $\bar{y} = \mathrm{Av}[y_i]$, and, $\bar{x} = \mathrm{Av}[x_i]$. The resulting mean-square error (MSE) is equal to,

$$\begin{aligned}\mathrm{MSE}(\hat{a}_0,\hat{a}_1) &= \sum_{i=1}^{n}(y_i-\hat{a}_1x_i-\hat{a}_0)^2 \\ &= \sum_{i=1}^{n}\left((y_i-\bar{y})-\hat{a}_1(x_i-\bar{x})\right)^2 \\ &= \mathrm{Av}\left[(y_i-\bar{y})^2\right]-\frac{\mathrm{Av}[(x_i-\bar{x})(y_i-\bar{y})]^2}{\mathrm{Av}[(x_i-\bar{x})^2]}.\end{aligned} \tag{22}$$

Moreover, for the specific model of measurements (10), and an asymptotically large number of measurements, $n \gg 1$, the LS estimate of a_1 can be rewritten as,

$$\hat{a}_1 = \frac{\mathrm{E}[\varepsilon_{xi}\varepsilon_{yi}]}{\mathrm{E}\left[\varepsilon_{xi}^2\right]} = \frac{\mathrm{cov}[\varepsilon_{xi}\varepsilon_{yi}]}{\mathrm{var}[\varepsilon_{xi}]}. \tag{23}$$

In this case, the resulting MSE is equal to,

$$\mathrm{MSE}(\hat{a}_0,\hat{a}_1) = \mathrm{E}\left[\varepsilon_{yi}^2\right]-\frac{\mathrm{E}[\varepsilon_{xi}\varepsilon_{yi}]^2}{\mathrm{E}\left[\varepsilon_{xi}^2\right]}. \tag{24}$$

Importantly, examining eqs. (22) and (24), it can be observed that the achievable MSE is greatly affected by the cross-covariance terms, $\mathrm{Av}[(x_i-\bar{x})(y_i-\bar{y})]^2$, and, $\mathrm{E}[\varepsilon_{xi}\varepsilon_{yi}]$, respectively. In practice, this cross-covariance can be expected to be much larger between the zero-mean processes representing the model inputs and outputs than between the measurement noises at the model inputs and outputs. Consequently, the LS estimation of the model parameters performs poorly when the input and output signals are noisy constants as assumed in (10). In such a case, some other strategy for identifying the model parameters has to be adopted.

A. Estimating the model inputs and outputs

In the absence of measurement noises, the $n = (p+1)$ measurements are sufficient to obtain the model parameters in (4) by inverting the matrix, $\boldsymbol{\Phi}$, i.e.,

$$\boldsymbol{a} = \boldsymbol{\Phi}^{-1}(Q(\boldsymbol{x}))Q(\boldsymbol{y}). \tag{25}$$

However, theoretical guarantees about the existence of the inverse, $\boldsymbol{\Phi}^{-1}$, are not considered further in this paper.

The noise in the measurements of the static model inputs and outputs can be suppressed statistically by taking repeated measurements. In particular, considering the input-output model (10), this leads to the problem of estimating the deterministic, but otherwise unknown constants in the zero-mean, stationary additive noises from multiple measurements.

Several strategies were proposed in the literature for estimating the deterministic (without any prior knowledge) parameters [12]. The minimum variance unbiased (MVUB), and among them, the best linear unbiased (BLUE) methods yield the estimators with the minimum variance, provided that they exist, and that they can be found. The LS estimator will perform poorly as argued in the previous subsection. The maximum-likelihood (ML) estimator is relatively easy to obtain for simple input-output signal models (10), and since it is asymptotically unbiased as well as consistent, this estimator is selected here. Furthermore, note that it is sufficient to only consider the estimators for one input-output signal, since all these input-output signals have the same model (10).

In particular, given n quantized measurements, x_i, $i = 1,2,\ldots,n$, the task is to derive an ML estimator of the constant, $\bar{x}$, in an additive noise, ε_{xi}. In this paper, we assume that the additive noise is zero-mean, Gaussian, and stationary with the variance, σ^2. If the measurements are unquantized, it is straightforward to show that the ML estimator is the arithmetic mean, i.e., [12]

$$\hat{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}x_i = \frac{1}{n}\sum_{i=1}^{n}(\bar{x}+\varepsilon_{xi}) = \bar{x}+\bar{\varepsilon}_{xi}. \tag{26}$$

In the case the measurements are quantized into integer values using the mapping (7), the probability of the measurement, $\check{x}_i = k$, where $k \in \mathbb{Z}$ can be computed as,

$$\Pr(\check{x}_i = k) = Q\left(\frac{\Delta(k-1/2)-\bar{x}}{\sigma}\right)-Q\left(\frac{\Delta(k+1/2)-\bar{x}}{\sigma}\right) \tag{27}$$

where the Q-function for the standard Gaussian variable is defined as,

$$Q(t) = \int_t^{\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-t^2/2}\,\mathrm{d}t. \tag{28}$$

Provided that the additive noise is also white, the measurements are independent, and the ML estimator maximizes the joint probability density,

$$\Pr(\{\check{x}_i\}_i) = \prod_{i=1}^{n}Q\left(\frac{\Delta(\check{x}_i-1/2)-\bar{x}}{\sigma}\right)-Q\left(\frac{\Delta(\check{x}_i+1/2)-\bar{x}}{\sigma}\right). \tag{29}$$

Taking the logarithm, and then the derivative by $\bar{x}$ (i.e., the parameter to be estimated), we obtain,

$$\begin{aligned}&\frac{\partial}{\partial\bar{x}}\log\Pr(\{\check{x}_i\}_i) = \\ &\qquad -\frac{1}{\sigma}\sum_{i=1}^{n}\frac{\dot{Q}\left(\frac{\Delta(\check{x}_i-1/2)-\bar{x}}{\sigma}\right)-\dot{Q}\left(\frac{\Delta(\check{x}_i+1/2)-\bar{x}}{\sigma}\right)}{Q\left(\frac{\Delta(\check{x}_i-1/2)-\bar{x}}{\sigma}\right)-Q\left(\frac{\Delta(\check{x}_i+1/2)-\bar{x}}{\sigma}\right)}.\end{aligned} \tag{30}$$

In order to find, for which value of $\bar{x}$, the expression (30) becomes zero to maximize the log-likelihood, we can linearize the Q-function and its derivative about the point, x_0, i.e.,

$$\begin{aligned}Q(x) &\approx Q(x_0)-\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x_0^2/2}(x-x_0) \\ \dot{Q}(x) &\approx \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x_0^2/2}(x_0x-x_0^2-1).\end{aligned} \tag{31}$$

The corresponding approximations are then,

$$\begin{aligned}Q(x_0-b)-Q(x_0+b) &\approx b\,\mathrm{e}^{-x_0^2/2}\sqrt{\frac{2}{\pi}} \\ \dot{Q}(x_0-b)-\dot{Q}(x_0+b) &\approx -x_0\,b\,\mathrm{e}^{-x_0^2/2}\sqrt{\frac{2}{\pi}}.\end{aligned} \tag{32}$$

Assuming $x_0 = (\Delta\check{x}_i-\bar{x})/\sigma$, and, $b = (\Delta/2)/\sigma$, in approximations (32), the derivative of the log-likelihood function (30) can be greatly simplified as,

$$\sum_{i=1}^{n}\frac{\Delta\check{x}_i-\bar{x}}{\sigma^2} \stackrel{!}{=} 0. \tag{33}$$

Consequently, we find that the ML estimator of $\bar{x}$, from the quantized noisy measurements, $\breve{x}_i$, is again a simple arithmetic average, i.e.,

$$\hat{\bar{x}} = \Delta \frac{1}{n} \sum_{i=1}^{n} \breve{x}_i. \tag{34}$$

However, and importantly, note that the ML estimator was derived under the assumption that, $b = (\Delta/2)/\sigma$, is relatively small (i.e., $b < 1$), so that the linearization is sufficiently accurate. The value, $\Delta/2$, also represents the maximum quantization error, and thus, we can conclude that the arithmetic average estimator can be expected to perform comparatively well as the arithmetic average estimator for the unquantized measurements, when $(\Delta/2) \ll \sigma$.

The similar derivation can be performed for the case of binary quantization (9) when the measurements are quantized to, -1, and, $+1$, values. Under the assumption that, $\bar{x} \ll \sigma$, the ML estimator (which, in this case, can be shown to be actually the MVUB estimator) becomes,

$$\hat{\bar{x}} = \sigma \sqrt{\frac{\pi}{2}} \frac{1}{n} \sum_{i=1}^{n} \breve{x}_i, \quad \breve{x}_i \in \{-1, +1\}. \tag{35}$$

Thus, the estimator for the binary quantized measurements requires knowledge of the noise standard deviation, σ.

IV. Estimator Variances

In this section, the goal is to compare the variances of the estimation errors for different estimators considered in the previous section. In particular, when the measurements are unquantized, the estimator (26) is unbiased, and its variance is simply,

$$\mathrm{E}\left[(\hat{\bar{x}} - \bar{x})^2\right] = \sigma^2/n. \tag{36}$$

When the measurements are uniformly quantized, the ML estimator (34) may be biased, i.e.,

$$\begin{aligned} \mathrm{E}\left[\hat{\bar{x}}\right] &= \frac{\Delta}{n} \sum_{i=1}^{n} \mathrm{E}[\breve{x}_i] = \frac{\Delta}{n} \sum_{i=1}^{n} \sum_{k=-\infty}^{\infty} k \Pr(\bar{x} = k) \\ &= \frac{\Delta^2}{\sqrt{2\pi}\sigma} \sum_{k=-\infty}^{\infty} k \, \mathrm{e}^{-\frac{(\Delta k - \bar{x})^2}{2\sigma^2}} \end{aligned} \tag{37}$$

where we assumed linearization (32) of the Q-function.

Further insight can be obtained by analyzing the best case, and the worst case quantization scenarios. In particular, without loss of generality, the best case scenario occurs, when $\bar{x} = 0$ (more precisely, if $\bar{x}$ is an integer multiple of Δ); then, the mean, $\mathrm{E}\left[\hat{\bar{x}}\right] = 0$, and the ML estimator (34) is unbiased. On the other hand, the largest bias occurs for the values, $\bar{x} = \pm\Delta/2$ (more precisely, if $\bar{x}$ is an odd-integer multiple of $\Delta/2$). Hence, let, $\bar{x} = -c\Delta/2$, where $c = 0$, represents the best case, and $c = 1$, represents the worst case scenario, respectively.

The ML estimator (34) with quantized measurements has the variance,

$$\begin{aligned} \mathrm{E}\left[\left(\hat{\bar{x}} - \mathrm{E}\left[\hat{\bar{x}}\right]\right)^2\right] &= \frac{\Delta^2}{n^2} \mathrm{E}\left[\sum_{i,j=1}^{n} \breve{x}_i \breve{x}_j\right] - \Delta^2 \mathrm{E}[\breve{x}_i]^2 \\ &= \frac{\Delta^2}{n} \left(\mathrm{E}\left[\breve{x}_i^2\right] - \mathrm{E}[\breve{x}_i]^2\right). \end{aligned} \tag{38}$$

To simplify the notation, define the moment [cf. (37)],

$$\begin{aligned} Z_m(\Delta/\sigma) &= \mathrm{E}[\breve{x}_i^m \mid \bar{x} = -c\Delta/2] \\ &= \sum_{k=-\infty}^{\infty} k^m \Pr(\breve{x} = k \mid \bar{x} = -c\Delta/2) \\ &= \frac{\Delta}{\sqrt{2\pi}\sigma} \sum_{k=-\infty}^{\infty} k \, \mathrm{e}^{-\frac{(k+c/2)^2}{2} \frac{\Delta^2}{\sigma^2}}. \end{aligned} \tag{39}$$

After substituting $Z_m(\Delta/\sigma)$ into (38), the final expression for the estimator variance becomes,

$$\mathrm{E}\left[\left(\hat{\bar{x}} - \mathrm{E}\left[\hat{\bar{x}}\right]\right)^2\right] = \frac{\Delta^2}{n} \left(Z_2(\Delta/\sigma) - Z_1^2(\Delta/\sigma)\right). \tag{40}$$

The derived MSE expression (40) is compared with the computer simulations in Figure 3 assuming $n = 100$ measurements, and the quantization intervals with $\Delta = 1/2$. It can be observed that the derived expression is in a good agreement with simulations, provided that the condition, $\Delta \ll \sigma$, is satisfied. For larger values of Δ/σ, the derived expression represents a loose lower bound of the actual MSE. As expected, when the estimator with quantized inputs is unbiased (the best case scenario), the MSE continues to be reduced by reducing the amount of measurement noise. When the quantization error makes the estimator to be biased, the MSE eventually saturates, as might be expected.

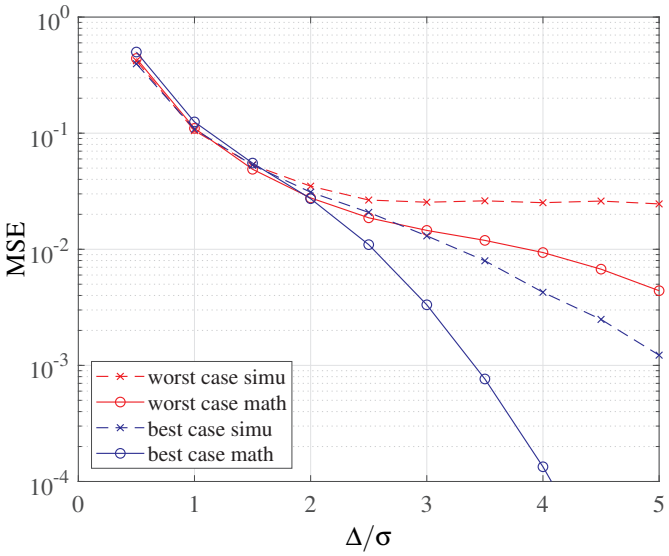


Figure 3. The MSE of the ML estimator with uniformly quantized measurements corresponding to the best case and the worst case quantization errors, respectively.

The variance of the MVUB estimator (35) with the binary quantized measurements can be shown to be,

$$\mathrm{E}\left[\left(\hat{\bar{x}} - \mathrm{E}\left[\hat{\bar{x}}\right]\right)^2\right] = \frac{\pi}{2} \frac{\sigma^2}{n}. \tag{41}$$

Thus, it is $(\pi/2)$ times larger than the variance (36) of the estimator from unquantized measurements, and importantly, provided that the condition, $\bar{x} \ll \sigma$ is satisfied.

The Cramér-Rao bound can be derived using again a linearization of the Q-function in the low signal-to-noise ratio (SNR) regime to obtain, [13]

$$\mathrm{E}\left[\left(\hat{\bar{x}}-\mathrm{E}\left[\hat{\bar{x}}\right]\right)^2\right] \geq J^{-1} = \frac{\sigma^2}{n}\frac{(1-Q(\bar{x}/\sigma))\,Q(\bar{x}/\sigma)}{\dot{Q}(\bar{x}/\sigma)} \qquad (42)$$

where J denotes the Fisher information matrix (a scalar value, here). The normalized Cramér-Rao bound, nJ^{-1}/σ^2, is shown in Figure 4 (black-line), together with the MSE of the estimator having the binary quantized measurements (41) (blue-line), and the MSE of the estimator with unquantized measurements (36) (red-line). It can be observed that the MSE raises quickly with improving SNR. In such a case, the binary quantization error starts dominating, and it cannot be reduced, for example, by simply increasing the number of measurements.

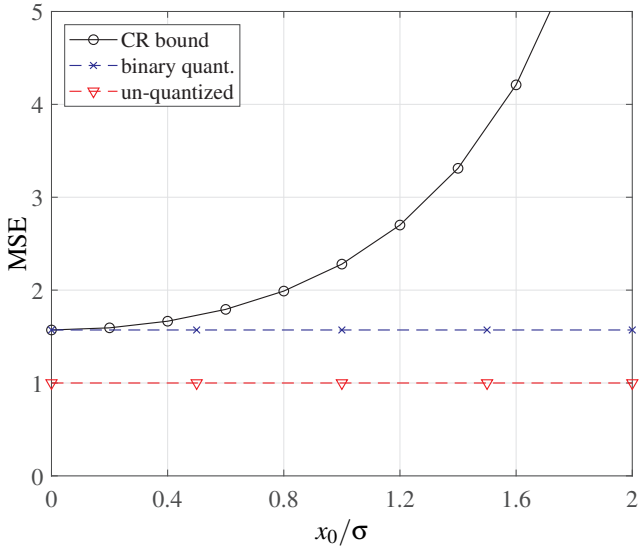


Figure 4. The Cramér-Rao bound of the estimator with the binary quantized measurements (black line), the actual MSE in the low-SNR regime (blue line), and the MSE of the estimator with the unquantized measurements (red-line).

V. Discussion and Future Work

Our investigations showed that the quantization noise can be neglected, provided that it is comparable with the measurement noise. If this condition is not satisfied, the estimators are only unbiased and consistent with respect to the additive measurement noise, and the estimation error is dominated by the residual quantization noise. The measurements obtained at both the system inputs and outputs represent a classical problem of system identification. When the inputs and outputs are static, i.e., they are constant values observed in an additive noise, the recommended strategy for estimating the model parameters is to first clean the input-output measurements by suppressing the measurement noises. This can be done independently for each input and output using different types of estimators. The noise-free input and output values can be then substituted into the model, and the model parameters are obtained by solving the same number of linear or non-linear equations representing the system model. This strategy is superior to classical least-squares model fitting (i.e., without suppressing the measurement noises first), provided that the inputs and outputs are noisy constant values. Furthermore, estimating the model parameters from input-output data pairs resembles a supervised machine learning. The main difference is that the data examples for machine learning are usually assumed to be noise-free, and the number of parameters assumed in machine learning models can be excessively large.

In this paper, our focus was on identifying relatively small linear models from their input-output measurements. Such models are common not only in engineering, but they also represent many physical laws. For example, Schrödinger and Maxwell's equations are both linear. It was shown in Figure 2 that the coarse-grained quantization can substantially affect the model, and also our perception of reality, if the model represents a physical law. This phenomenon is referred to here as Minecraft of system modeling, since the quantization makes the reality to appear as if it consisted of discrete blocks.

The future work can investigate the optimum representations of MIMO systems with discretized inputs and outputs. The non-linear systems can be modeled by recursive structures [14]. The fundamental question is how to suppress the quantization noise akin to suppressing the measurement noise. In this paper, the static input and output values were considered. Measuring the systems having the random processes as their inputs and outputs is more challenging, as it requires precise time-synchronization of the measurements at all the system inputs and outputs.

References

[1] P. Loskot, "Key ideas in parameter estimation," in *Proc. SIGNAL*, 2004, pp. 19–27.
[2] S. Haykin, *Adaptive Filter Theory*, 5th ed. Pearson Education, Harlow, England, 2014.
[3] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. Springer-Verlag New York, USA, 2000.
[4] J. C. Keenan and J. B. Lewis, "Estimation with quantized measurements," in *IEEE ICDC*, 1976, pp. 1284–1291.
[5] M. Fu and C. E. de Souza, "State estimation for linear discrete-time systems using quantized measurements," *Automatica*, vol. 45, no. 12, pp. 2937–2945, Dec. 2009.
[6] H. Poor, "Fine quantization in signal detection and estimation," *IEEE Trans. Info. Theory*, vol. 34, no. 5, pp. 960–972, Sep. 1988.
[7] S. Vardeman and C.-S. Lee, "Likelihood-based statistical estimation from quantized data," *IEEE Trans. Instrum. Measurements*, vol. 54, no. 1, pp. 409–414, Feb. 2005.
[8] A. Arafa, K. Banawan, K. G. Seddik, and H. V. Poor, "Sample, quantize, and encode: Timely estimation over noisy channels," *IEEE Trans. Communications*, vol. 69, no. 10, pp. 6485–6499, Oct. 2021.
[9] P. Stoica, X. Shang, and Y. Cheng, "The cramér-rao bound for signal parameter estimation from quantized data," *IEEE Sig. Proces. Magazine*, vol. 39, no. 1, pp. 118–125, Dec. 2021.
[10] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. Springer-Verlag, New York, USA, 1994.
[11] P. Loskot, "Polynomial representations of high-dimensional observations of random processes," *Mathematics*, vol. 9, no. 123, pp. 1–24, Jan. 2021.
[12] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, Upper Saddle River, NJ, USA, 1993, vol. I.
[13] Z. Hrdina, *Statistical Radio-Engineering*. Publishing Company of the Czech Technical University of Prague, 1996, in Czech.
[14] S. S.-T. Yau, X. Chen, X. Jiao, J. Kang, Z. Sun, and Y. Tao, *Principles of Nonlinear Filtering Theory*. Springer Nature Switzerland AG, 2024.