IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received May 6, 2021, accepted May 18, 2021, date of publication May 27, 2021, date of current version June 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3084115

Symbol-Level Synchronization Channel Modeling With Real-World Application: From Davey-MacKay, Fritchman to Markov

SHAMIN ACHARI, DANIEL G. HOLMES, AND LING CHENG, (Senior Member, IEEE)
School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg 2000, South Africa

Corresponding author: Ling Cheng (ling.cheng@wits.ac.za)

This work was supported in part by the South African National Research Foundation under Grant 129311 and Grant 132651.

ABSTRACT Errors in realistic channels contain not only substitution errors, but synchronization errors as well. Moreover, these errors are rarely statistically independent in nature. By extending on the idea of the Fritchman channel model, a novel error category-based methodology for determining channel characteristics is described for memory channels that contain insertion, deletion, and substitution errors. The practicality of such a methodology is reinforced by making use of real communication data from a visible light communication system. Simulation results show that the error-free and error runs using this new method of defining the channel clearly deviates from the Davey-MacKay synchronization model which is memoryless in nature. This further emphasizes the inherent memory in these synchronization channels which we are now able to characterize. Additionally, a new method to determine the parameters of a synchronization memory channel using the Levenshtein distance metric is detailed. This method of channel modeling allows for more realistic communication models to be simulated and can easily extend to other areas of research such as DNA barcoding in the medical domain.

INDEX TERMS Channel modeling, finite-state Markov channel, memory models, synchronization models.

I. INTRODUCTION

Systems which exhibit a correlation between errors while also having synchronization issues are quite common in practical, real-life applications. Thus, a method to characterize and model such systems proves beneficial. A few of these applications involve data transmission especially in cases where the channel is significantly harsh. Visible Light Communication (VLC) and Free Space Optical (FSO) communications are examples of such channels because they suffer drastically under the influence of interference, signal blocking, and turbulence. The applications are not only restricted to the domain of telecommunications either and can easily be extended to domains such as medicine. An example of this is described in [1] and [2] where Kracht and Schober modify the idea of watermark codes and synchronization error channels described by Davey and MacKay [3] to model and correct for errors while barcoding DNA in the process of DNA sequencing. Kracht et al describes how the system works well, but would benefit from a more complex model which incorporates memory into the channel as DNA sequencing channels are known to exhibit correlations between errors.

There are numerous channel models which take into account a combination of substitution errors as well as insertion and deletion errors (commonly referred to as synchronization errors). The most cited of which include the Gallager channel model [4], Zigangirov channel model [5] and the Davey-MacKay (DM) channel model [3], [6]. In terms of memory channels, an extensive review is conducted on the relevant error control techniques and is presented in [7]. More recently, [8] provides an in-depth look at the modeling of FSMC for fading channels. There, however, appears to be little mentioned regarding synchronization memory channels. The focus of this paper will remain on generative channel models where, in this case, the most cited include the Gilbert channel model [9], the Gilbert-Elliott (GE) channel model [10] and the Fritchman channel model [11].

Channel models currently exist for systems with discrete synchronization errors, and separately for those which are capable of characterizing memory. However, to the knowledge of the authors, no such model or modeling technique exists where both scenarios are taken into account in a single

The associate editor coordinating the review of this manuscript and approving it for publication was Marco Martalo.

model. In this paper, we provide a methodology with relevant channel models to analyze systems with correlated errors in the presence of insertion and deletion errors. The first contribution of this paper is to introduce the idea of using the Fritchman Model and consequently Hidden Markov Models, which inherently contain memory to model substitution errors in addition to synchronization errors by making use of various error groupings. The second contribution presented in this paper is a novel Finite-State Makov Channel (FSMC) model which contains states for insertions, deletions, substitutions and transmission. This novel FSMC provides a more comprehensive model for real-world scenarios and the applicability of such a model is reinforced by using communication data from an actual VLC testbed. The use of real-world data in this analysis also corroborates our intuitive notions of when we could possibly experience correlated synchronization errors and indicates when the presented models may be beneficially used.

The rest of the paper is structured as follows. The DM channel, as well as the Fritchman model, are further detailed in Section II along with the metrics used for analysis. This is followed by Section III where the system setup and approach is described. Section IV then shows an analysis using a modified Fritchman model to characterize synchronization errors. From this, a novel channel that consists of both synchronization errors and memory is formulated and discussed in Section V. Conclusions are drawn in Section VI.

II. BACKGROUND

A. DAVEY-MACKAY SYNCHRONIZATION CHANNEL MODEL

The Gallager, Zigangirov and DM models are all binary, discrete, and memoryless, which means they tend towards an independent and identically distributed (IID) classification. Additionally, none of these channels are able to indicate the positions of errors. For this paper, we focus on the DM channel as it is the most comprehensive and incorporates elements from both the Gallager and Zigangirov models. In fact, Leigh shows that both the Zigangirov and DM channels are equivalent when the parameters are specifically defined [12]. More comprehensive details regarding the Gallager channel model are found in [4], [12], [13] and likewise further information regarding the Zigangirov channel is available in [5], [12], [13].

In the DM synchronization channel, the queued bits awaiting transmission may undergo one of three events to proceed to the next time step. Bits may be inserted into the received stream with a probability of P_i. Since there is, in theory, an unlimited number of possible insertions, for n insertions the probability is given as P_i^n. A transmission or deletion must follow an insertion to allow the system to move into the next time step. A maximum number of insertions I is imposed on the system for simplification. With a probability P_d, a bit is deleted from the stream and does not appear in the received sequence. Lastly, with a probability P_t, a bit is transmitted where $P_t = 1 - P_i - P_d$ [3], [6], [12]. Since bit flips are accounted for, a substitution error may occur on a transmitted bit with a probability P_s. As such, the sum of the probabilities of a substitution and no substitution must equal unity. The DM channel model is better illustrated in Figure 1 [13] where t_i and t_{i+1} indicates the time steps at time i and $i+1$ respectively.

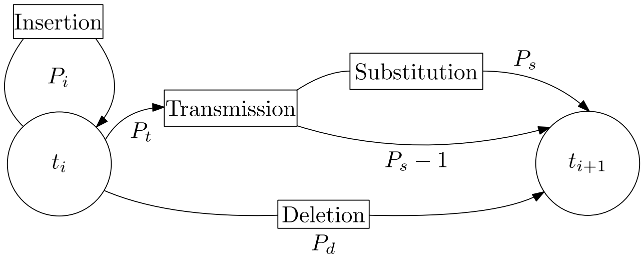


FIGURE 1. Davey-MacKay synchronization channel model.

Determining the closed-form expression for the capacity of an Insertion, Deletion, and Substitution (IDS) channel is quite complex and in fact, it still remains an open problem in the field of communication and synchronization models. As such, [3] and [6] provide empirical capacities for the DM channel for various parameters.

B. FRITCHMAN CHANNEL MODEL AND PARAMETER ESTIMATION

The simplest way of describing how errors occur in a channel is with the use of a discrete, memoryless channel, where the current output is only dependent on the current input [14]. Given an input alphabet $X = x_0, x_1, \ldots, x_{q-1}$ and an output alphabet $Y = y_0, y_1, \ldots, y_{Q-1}$, then the set of $q \times Q$ conditional probabilities are given by $P(Y = y_i|X = x_i) = P(y_i|x_i)$. These parameters are able to completely define a discrete, memoryless channel [14], [15]. Unfortunately, most real channels exhibit some memory within the system, where the cause of one error tends to create more errors within that region of transmission [14]. An easy method of overcoming this memory in substitution error channels is by converting them into memoryless channels with the use of interleavers, as this spreads the errors throughout the sequence making the distribution IID in nature and "locally memoryless" [16]. While the process of interleaving simplifies the modeling it adds additional system complexity and delays while also foregoing the additional channel capacity we may gain by utilizing the channels inherent memory [16]. Additionally, the use of interleavers poses significant problems when dealing with synchronization errors, as there is no accurate way of knowing how many and in what positions the bits were inserted and deleted. This makes it near impossible to determine the depth of the interleaver required and thus drastically restricts the use of them in these considered channels.

Since the models of interest are generative, we are able to generate statistics based on the error sequences. The alternate category of discrete channel models are descriptive and are

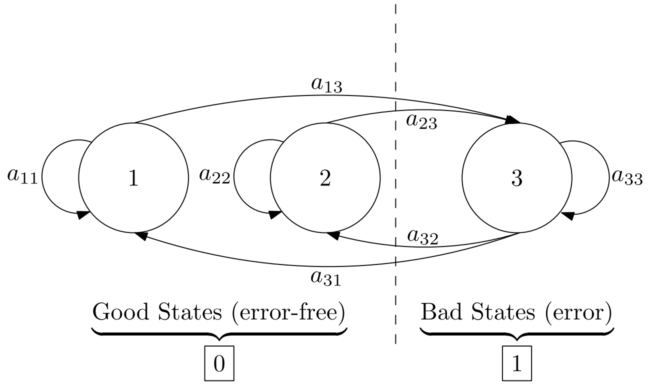


FIGURE 2. **Three state Fritchman model with single error state.**

discussed further in [14]. The Gilbert, GE, and Fritchman models all have the ability to model memory in a channel where the errors are binary in nature and have some statistical dependence between them [14]. Additionally, all these models employ the use of finite-state Markov models [14], [17], [18]. This paper once again limits the discussion to Fritchman models for statistically dependent error channels as it is the most comprehensive. It has also gained substantial attention in recent years due to the practicality it offers in modeling realistic communication channels and the ease of parameter estimation [17]. Readers may find more information of the Gilbert and GE channel models in [9] and [10] respectively.

Fritchman used a finite state partitioned Markov model to model binary errors where the partitioning was done according to error-free and error states. The Fritchman model contains N total states, of which K states are partitioned as good states and the remaining $N-K$ states are bad states. A condition is also applied where a good state is error-free and must produce a 0 in the error sequence, whereas a bad state is erroneous in nature and will always produce a 1 in the error sequence [11], [14]. Using a single error state not only reduces complexity, but it allows the model parameters to be uniquely specified. It also reduces the model parameters from $2K(N-K)$ parameters to $2(N-1)$ parameters. In a single error state model, the error free run distribution can completely specify the model parameters [11]. Figure 2 shows a simplified 3 state Fritchman model with a single error state, along with the transition and emission matrices shown in Equations (1) and (2) respectively [11], [17], [18].

$$A = \begin{bmatrix} a_{11} & 0 & a_{13} \\ 0 & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{1}$$

$$B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

The Fritchman model makes use of an empirical technique where the exponentials from (A.3) and (A.4) are used to fit the measured data [17]. Since the Fritchman model is designed on Markov processes, an easier method of parameter estimation is based on the iterative procedural Baum-Welch algorithm. This algorithm converges to a maximum likelihood estimator of $\{A, B\}$ which seeks to maximize $Pr(O|\{A, B\})$ [17]. In other words, the Baum Welch algorithm is used to find the most likely transition and emission matrices that could produce a set of observations. The Baum-Welch algorithm is detailed in [17]. It is worth noting that for the Fritchman model, only the transition matrix, A, is estimated as the entries for the emission matrix, B, are fully known. Additionally, O corresponds to the observed sequence and a_{ij} is the probability of transitioning from state i to state j (the respective entry in the i^{th} row and j^{th} column of the transition matrix). The Baum-Welch algorithm is used in this manner to estimate the transition matrices for all relevant simulations in this paper.

Determining the channel capacity of the Fritchman model is somewhat of an easier task than that of a memoryless IDS channel. [19] provides a closed-form equation for the average entropy of a stationary, ergodic Markov process and is shown in Equation (3). Here H_i corresponds to the entropy of state i and is calculated using Equation (4). ρ_i is the probability that the source is in state i or in other words the stationary state probability or steady-state probability of been in state i. From this, the capacity of the model, C, is readily calculated from Equation (5).

$$\overline{H} = \sum_{i=1}^{N} \rho_i H_i \quad \textit{bits/symbol} \tag{3}$$

$$H_i = -\sum_{j=1}^{N} a_{ij} \log_2(a_{ij}) \quad \textit{bits/symbol} \tag{4}$$

$$C = 1 - \overline{H} \quad \textit{bits/symbol} \tag{5}$$

C. PERFORMANCE METRICS FOR MODEL ANALYSIS

To effectively analyze the proposed system, we use the Chi-Squared (χ^2) Goodness of fit test as well as the Mean Squared Error (MSE). The χ^2 and MSE values can be calculated using Equations (A.1) and (A.2) respectively. These statistics will determine how well the observed data fits with the expected data. We then look at error-free and error run distributions to characterize the channel models. The χ^2, MSE and procedure used for the tests are outlined in Appendix A along with a brief description of the error-free and error run distributions.

III. ERROR CATEGORY-BASED CHANNEL MODELS SETUP

As mentioned previously, the applications for this type of analysis is widespread. To better solidify the practicality of this methodology, data collected from a VLC testbed is used in this approach. The data is publicly available from [20] and is originally used in [21], which describes an inter and intra-vehicle data communication system based on VLC. We limit the parameters of the VLC data to use 1 synchronization word and a frame length of 10003 symbols to simplify the analysis, but this procedure can easily be extended to other parameters and data sets. In particular, we look at low SNR, low baud rate communication (1.32dB at

50 Kilobits per second), as well as high SNR, high baud rate communication (18dB at 1 Megabit per second), as this is where the most errors in the system occur without complete failure. This procedure focuses on the symbol level (bit-level for binary systems) as the transitions within a synchronized frame are analyzed.

In this approach, we assume that the receiver has full knowledge of the transmitted data and as a result, the Levenshtein Distance (edit distance) algorithm may be used to determine the most likely states (insertion, deletion, substitution or transmission) that the channel traverses during communication [22]–[24]. This state path is hereon referred to as the synchronization error sequence and it is obtained for various communication parameters. Additionally, the error probabilities for P_t, P_s, P_i and P_d are calculated from this synchronization error sequence by summing the number of occurrences of a certain state and dividing this by the total length of the sequence. These error probabilities allow us to simulate communication over the DM synchronization channel for comparison. Depending on the type of error category of interest, we can also calculate the overall error probability, P_e, which will be used to simulate the IID plots.

An example synchronization error sequence for an IDS channel could be, **t,t,s,t,t,t,t,i,t,t,t,t,d,t,t,t,t,t,s,t** where **t** describes an error-free transmission, an **s** represents a bit flip or substitution error and an **i** and **d** represents an insertion and deletion, respectively.

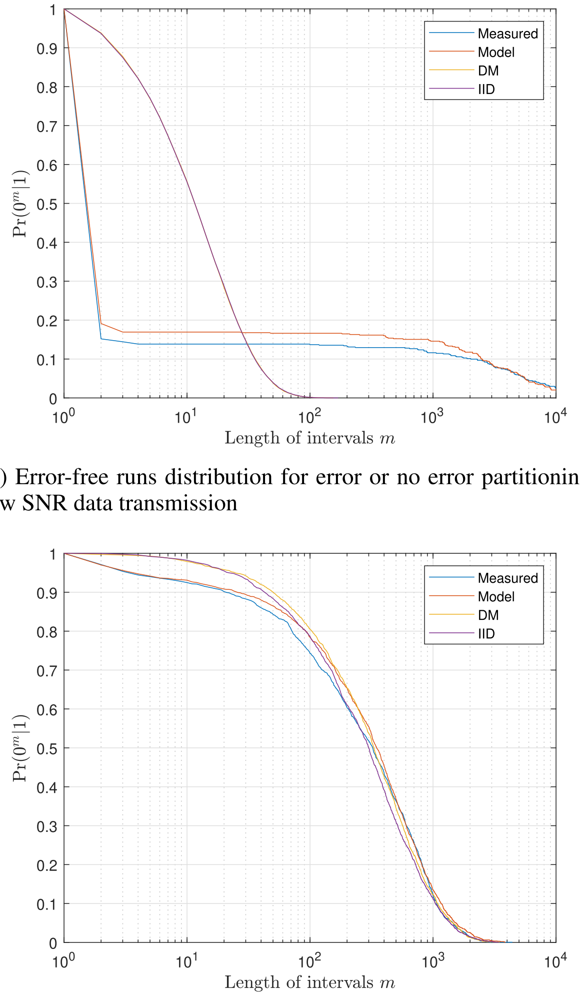
The next step is to generate the channel model. For simplicity, we limit the Fritchman model to three states, with a single error state, as there tends to only be slight accuracy gains with much more complexity for the additional states [25]. Using this approach will require some manipulation as the model is binary in nature, and a synchronization channel produces a variety of errors (insertions, deletions and substitutions). As such, the procedure will convert the synchronization channel errors encountered in an error sequence into a binary error sequence. This will be used to create a channel model based on the Fritchman model and Baum-Welch algorithm to ultimately determine the parameters of the channel.

To convert this synchronization error sequence to a binary form, we will look at five different categorisations or error category-based groupings: Error or Error-Free, Synchronization Error or No Synchronization Error, Substitution Error or No Substitution Error, Insertion Error or No Insertion Error, and lastly Deletion Error or No Deletion Error. These are further explained in the next section.

IV. ERROR CATEGORY-BASED CHANNEL MODELS ANALYSIS

A. ERROR CATEGORY 1: ERROR OR ERROR-FREE

Firstly, the synchronization error sequence could be converted into a binary error sequence by looking for either an error or no error. In this case, all errors encountered (insertion, deletion, and substitution) in the synchronization error sequence will be classified as an error and produce a **1** in the binary error sequence, whereas no error is a perfect transmission and produces a **0** in the binary error sequence. Using the example synchronization error sequence from the previous section will produce **0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,1,0** as the binary error sequence for this case. Figure 3 shows the error-free run distributions for the measured VLC system data (Measured), the corresponding Fritchman model simulated data (Model), the DM model simulated data, and lastly an IID sequence.

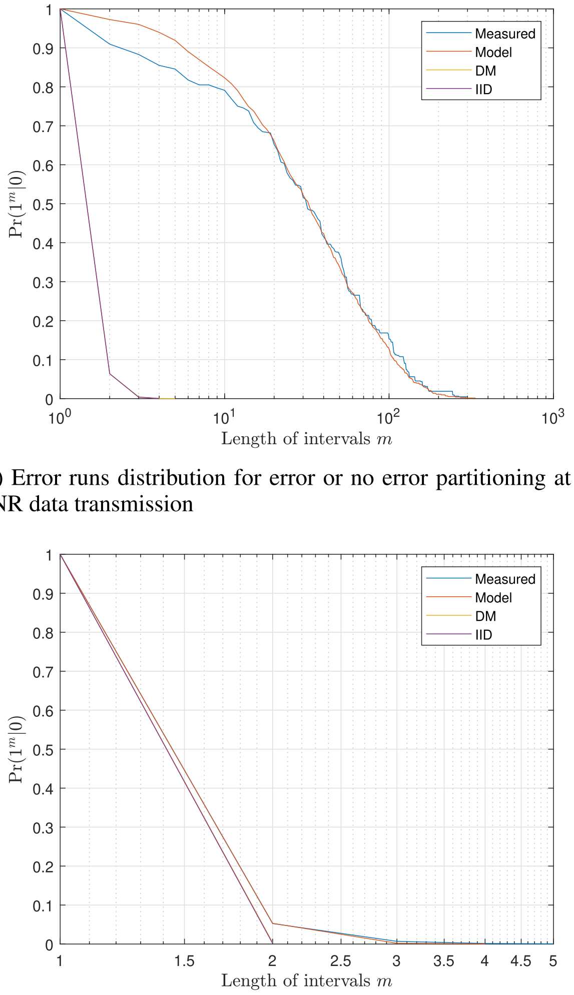


(a) Error-free runs distribution for error or no error partitioning at low SNR data transmission

(b) Error-free runs distribution for error or no error partitioning at high SNR data transmission

FIGURE 3. Error-free run distributions of Measured VLC data, simulated Model data, simulated DM model data and IID data when partitioning error sequence according to Error or no Error.

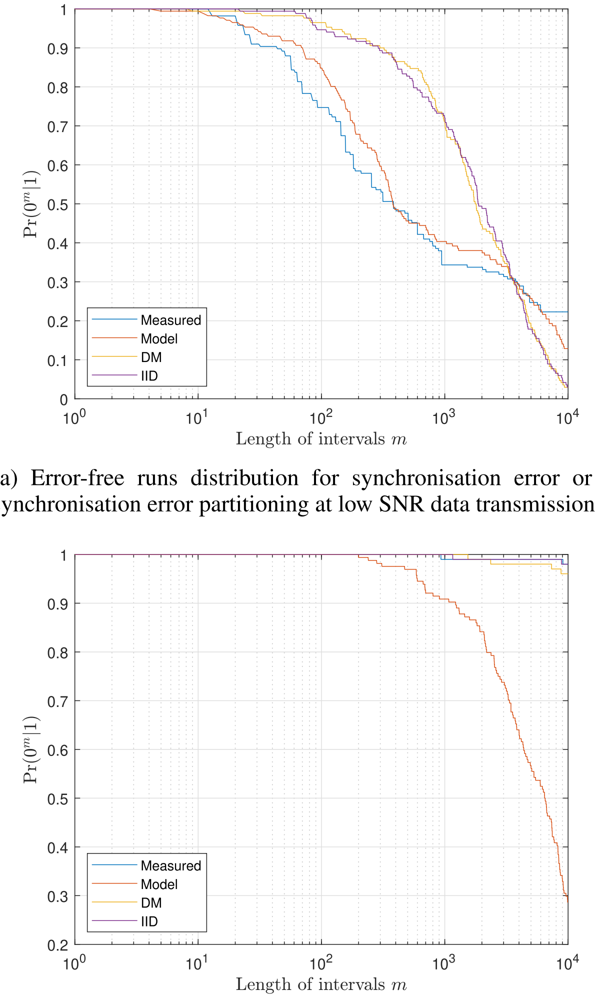
It is evident from Figure 3a, which shows communication at low SNR, that the observed channel (Measured) contains a significant amount of memory as it deviates substantially from the IID plot which is by definition memoryless. The Model plot is almost identical to the measured data, which shows the Fritchman model and the parameters generated accurately depict the real channel. Additionally, the DM plot closely follows the IID plot which reiterates the idea that there



(a) Error runs distribution for error or no error partitioning at low SNR data transmission

(b) Error runs distribution for error or no error partitioning at high SNR data transmission

FIGURE 4. Error run distributions of measured VLC data, simulated model data, simulated Davey-Mackay model data and IID data when partitioning sequence according to Error or no Error.



(a) Error-free runs distribution for synchronisation error or no synchronisation error partitioning at low SNR data transmission

(b) Error-free runs distribution for synchronisation error or no synchronisation error partitioning at high SNR data transmission

FIGURE 5. Error-free run distributions of measured VLC data, simulated model data, simulated Davey-Mackay model data and IID data when partitioning sequence according to synchronization error or no synchronization error.

is no memory between errors within the DM synchronization model. Communication at higher SNR, shown in Figure 3b, shows that all the plots are similar, especially after run lengths of around 100 consecutive error-free transmissions. This follows intuition as there are less errors encountered at higher SNR values. Thus, the few errors produced during transmission would likely be more sporadic.

Looking at Figure 4, which shows the error-run distribution of the same models and parameters as above, similar trends are observed. In both the low and high SNR scenarios, the DM channel data closely follows that of the IID, while the Measured and Model data are highly correlated but deviate from the IID. The deviation is more significant in the case of the low SNR, shown in Figure 4a. While there is still a difference between the measured data and IID in the high SNR, shown in Figure 4b, it is almost negligible. We once again confirm the accuracy of the generated model as it accurately fits the measured data. It is also shown in the low SNR case that once an error is experienced, it is likely to cause another error. Variable cluster sizes are seen, sometimes exceeding over 100 consecutive erroneous digits. For the IID and DM data, the cluster of consecutive errors rarely exceeds 5 bits or symbols for the low SNR case. This is because there is no form of memory, and it is highly unlikely to see many consecutive errors.

B. ERROR CATEGORY 2: SYNCHRONIZATION ERROR OR NO SYNCHRONIZATION ERROR

Next, only synchronization errors are isolated. In this case, a transmission and substitution error will produce a **0** in the binary error sequence stream, whereas an insertion or deletion produces a **1**. Using the example synchronization error sequence in Section III will produce **0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,1** as the binary error sequence for this case. The error-free run distribution for the four plots with this partitioning of errors is shown in Figure 5.

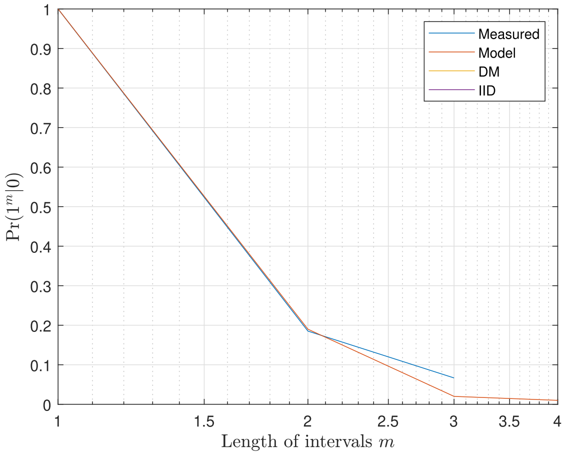


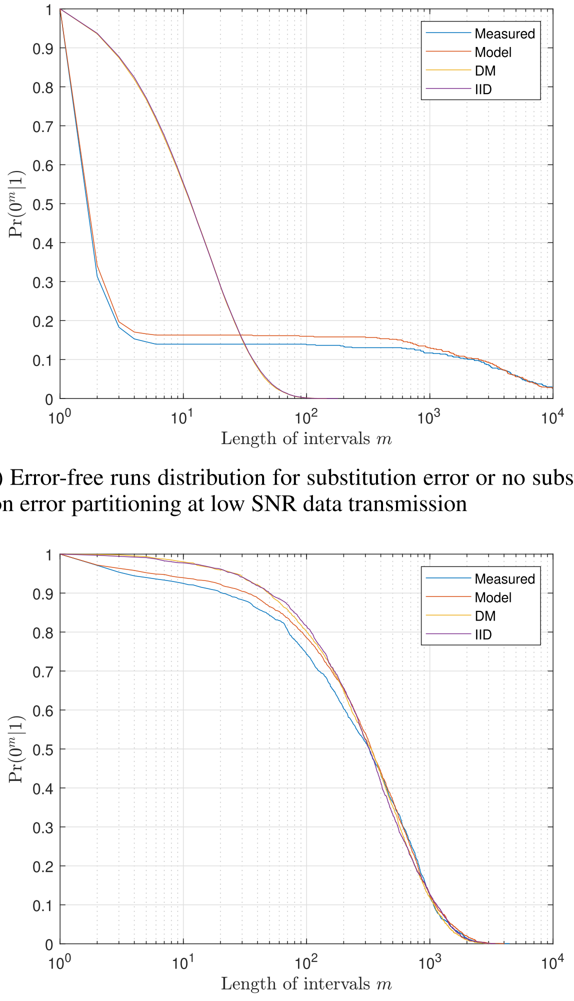
FIGURE 6. **Error run distributions of measured VLC data and simulated model data when partitioning sequence according to synchronization error or no synchronization error at low SNR data transmission.**

It is evident in the low SNR case, shown in Figure 5a, that the DM model follows an IID trajectory quite closely while the Measured data, and subsequently the Fritchman Model data, deviates from it. This, again, indicates memory between symbols of correct transmission and synchronization errors. The Model plot closely follows the Measured data, showing that the model created using the above process adequately describes our system characteristics. The plots in Figure 5b are inconclusive as there were not enough synchronization errors present at high SNR values to be accurately modelled and simulated.

Figure 6 shows the error run distributions for the Measured VLC system data and the corresponding models simulated data for the synchronization no synchronization error partitioning at low SNR. It can be seen that there are at most two consecutive synchronization errors for this system, and the occurrence of synchronization errors, in general, are quite low. For this reason, the plots of the IID and DM channel are not visible, as simulating these channels with such low error probability allowed for at most a single synchronization error between error-free runs. However, it is worth noting that while this system may not have substantial synchronization errors, the procedure and methodology used can still be applied for harsher channels where more severe synchronization errors do exist.

C. ERROR CATEGORY 3: SUBSTITUTION ERROR OR NO SUBSTITUTION ERROR

In this error category, we focus on how substitution errors affect the error-free and error run distributions. In this scenario, a transmission and synchronization error will produce a **0** in the binary error sequence stream, whereas a substitution error produces a **1**. Using the example synchronization error sequence from Section III will produce **0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0** as the binary error sequence for this case. The error-free run distribution for the various plots with this partitioning of errors is shown in Figure 7, where 7a and 7b show the plots for low SNR and high SNR communication respectively. These error-free and error run distribution plots are almost identical to the plots shown in Figure 3 and Figure 4. This indicates that the most common type of error encountered in the system are substitution errors and a similar insight into them naturally follows.



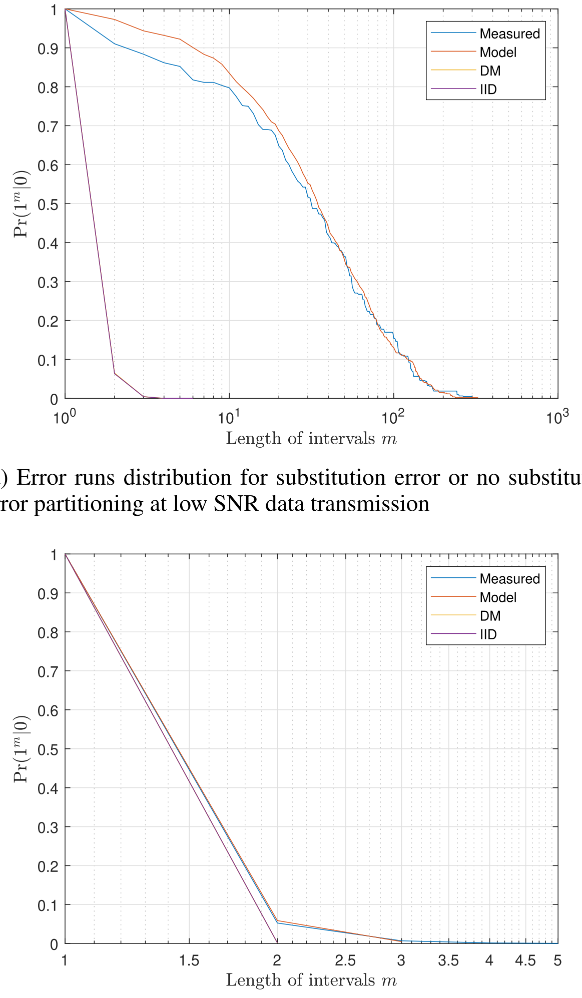
(a) Error-free runs distribution for substitution error or no substitution error partitioning at low SNR data transmission

(b) Error-free runs distribution for substitution error or no substitution error partitioning at high SNR data transmission

FIGURE 7. **Error-free run distributions of measured VLC data, simulated model data, simulated Davey-Mackay model data and IID data when partitioning sequence according to substitution error or no substitution error.**

D. ERROR CATEGORY 4: INSERTION ERROR OR NO INSERTION ERROR

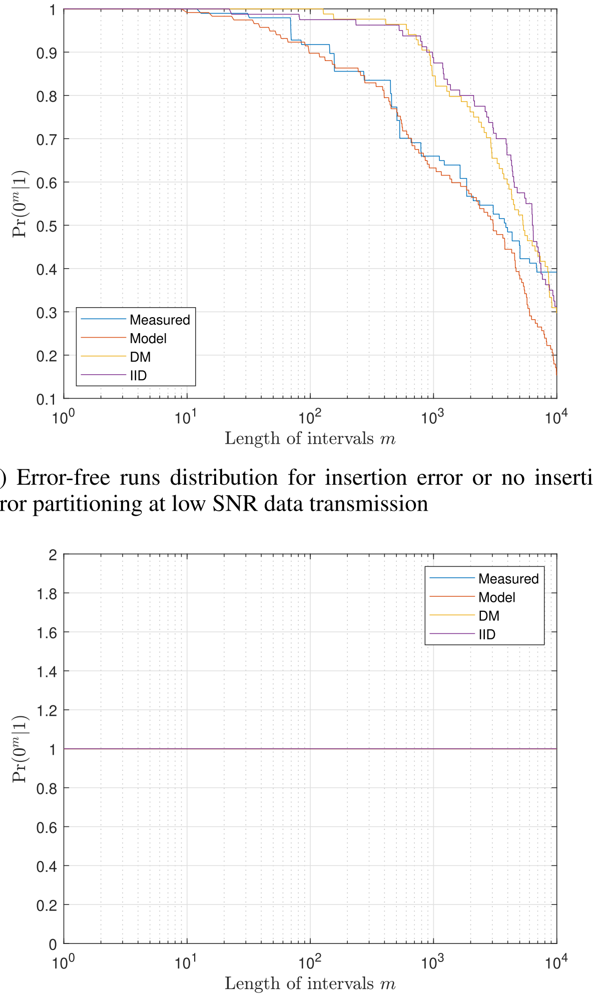
The following case is added for completeness of the analysis as the previous error categories already indicated that there is a low probability of producing synchronization errors in the system, and subsequently insertion errors. For the case where partitioning is done according to insertion or no insertion error, the plots for the error-free runs and error runs are shown in Figure 9 and Figure 10, where each subfigure shows



(a) Error runs distribution for substitution error or no substitution error partitioning at low SNR data transmission

(b) Error runs distribution for substitution error or no substitution error partitioning at high SNR data transmission

FIGURE 8. Error run distributions of measured VLC data, simulated model data, simulated Davey-Mackay model data and IID data when partitioning sequence according to substitution error or no substitution error.



(a) Error-free runs distribution for insertion error or no insertion error partitioning at low SNR data transmission

(b) Error-free runs distribution for insertion error or no insertion error partitioning at high SNR data transmission

FIGURE 9. Error run distributions of measured VLC data, simulated model data, simulated Davey-Mackay model data and IID data when partitioning sequence according to insertion error or no insertion error.

low and high SNR data transmission. Only transmission for low SNR is shown for the error run distributions as there are too few insertions that occur at sufficiently high SNR communication. Since our interest for this error category is in the analysis of insertion errors, an insertion will produce a **1** in the binary error sequence, while a transmission, substitution and deletion will all be analyzed as no error and produce a **0**. For the given synchronization error sequence this will produce a binary error sequence of **0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0**.

E. ERROR CATEGORY 5: DELETION ERROR OR NO DELETION ERROR

Once again, the following case is added for completeness for the reasons explained above. For this case, where partitioning is done according to a deletion or no deletion error, the binary error sequence obtained from the example synchronization error sequence is **0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0**. To obtain this sequence, a **1** is placed in the binary error sequence when a deletion occurs in the synchronization error sequence. All other likely states produce a **0** in the binary error sequence. The plots for the error-free runs and error runs are shown in Figure 11 and Figure 12. Again, only the error-free runs have an associated low and high SNR communication. The error run distribution only indicates the low SNR communication as too few deletion errors are produced at high SNR communication. These plots once again are almost the same as the plots obtained for the cases of synchronization error or no synchronization error, and insertion error or no insertion error. As such, the analysis and observations would follow similar explanations.

F. CHI-SQUARED AND MSE ANALYSIS OF PLOTS

Tables 1 - 5 outline the various Chi-Squared and MSE values obtained for various plots under different communication parameters. To better explain the values found in each table,

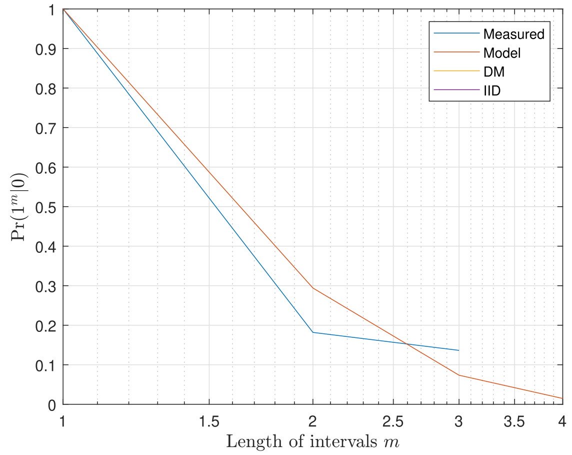
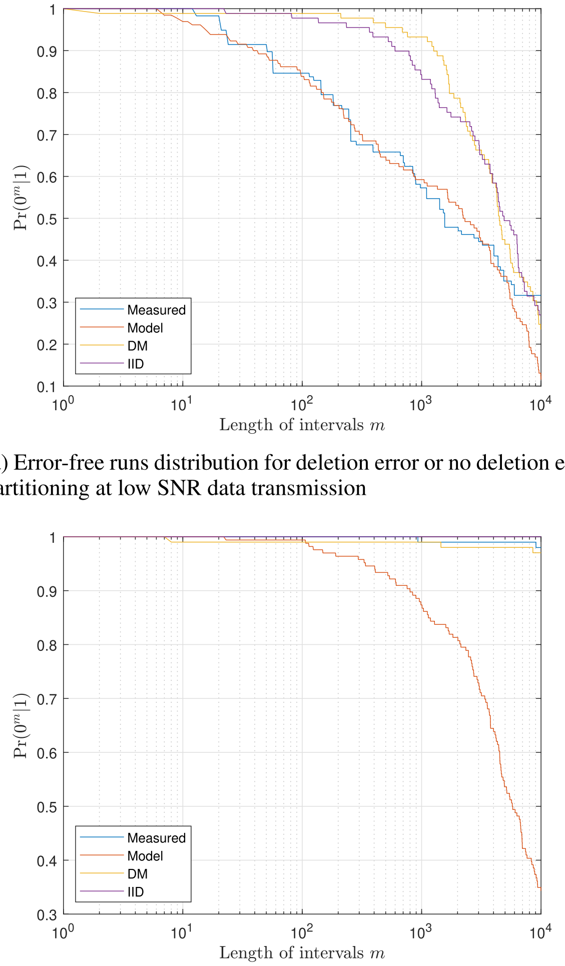


FIGURE 10. Error run distributions of measured VLC data and simulated model data when partitioning sequence according to insertion error or no insertion error at low SNR data transmission.



(a) Error-free runs distribution for deletion error or no deletion error partitioning at low SNR data transmission

(b) Error-free runs distribution for deletion error or no deletion error partitioning at high SNR data transmission

FIGURE 11. Error run distributions of measured VLC data, simulated model data, simulated Davey-Mackay model data and IID data when partitioning sequence according to deletion error or no deletion error.

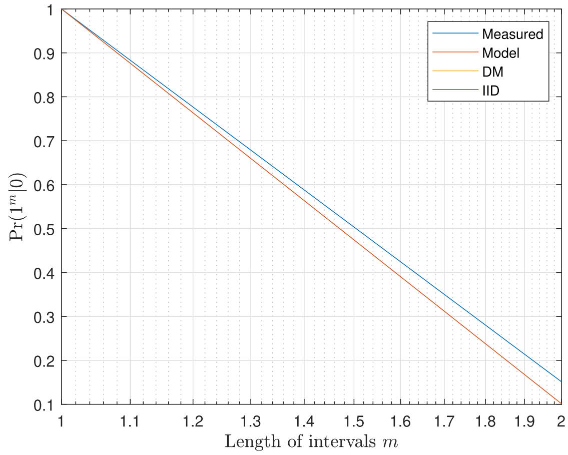


FIGURE 12. Error run distributions of measured VLC data and simulated model data when partitioning sequence according to deletion error or no deletion error at low SNR data transmission.

we will make use of the entries in the first block of Table 1. Here the values correspond to low SNR communication when the bin width is set to 1 and the plots compared are the IID and the Measured data when using the Error or Error-Free categorisation. As can be read off the table, a χ^2 value of 7202 is obtained which leads to a p-value of 0 as the df (degrees of freedom) is 35. This leads to the conclusion that the null hypothesis, H_N should be rejected as the p-value is less than the 0.01 significance value that is used. This implies that there is a significant difference between the IID and Measured data plots for the given parameters, which again reiterates the notion that the measured VLC data contains memory as it deviates significantly from a memoryless IID. Likewise, the MSE column indicates an MSE value of 8113 which is calculated using $k = 36$ data points. Here, the lower the relative MSE value, the more alike the compared plots are. While these values vary depending on the bin widths used (this may need to be optimised for a true representation), they still provide a good general indicator of how similar or contrasting the plots are to each other. In particular, it is noticed that at low SNR communication, the Model data and Measured data are generally in agreement with each other while significantly differing from the DM and IID channel plots. It is also evident that at these low SNR communication parameters, the DM plots are quite similar to an IID distribution. At high SNR communication, all the plots start to converge and we see a decrease in the Chi-Square and MSE values for these parameters. From the analysis of the real-world data, it is evident that our notions of when we will experience correlated errors are correct. At high SNR communication, the channel experiences fewer errors and starts to degenerate into an IID model. Again, we refer the reader to Appendix A for a discussion on these metrics as well as the method implemented for the calculation of each statistic.

V. A NOVEL MEMORY SYNCHRONIZATION CHANNEL MODEL

While the above setup and analysis is useful for modeling memory in synchronization channels, it is not without

TABLE 1. Values obtained for Chi-Squared and MSE analysis for various plots and bin widths using the Error or Error-Free categorisation.

	Low SNR						**High SNR**					
	Bin Width = 1		Bin Width = 5		Bin Width = 10		Bin Width = 1		Bin Width = 5		Bin Width = 10	
	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE
IID vs Measured	7202 (p-value =0) df = 35 Verdict: Reject H_N	8113 $k = 36$	1303 (p-value =1.284e-271) df = 12 Verdict: Reject H_N	14802 $k = 13$	550 (p-value =1.485e-115) df = 6 Verdict: Reject H_N	14977 $k = 7$	6 (p-value =0.01669) df = 1 Verdict: Accept H_N	6385 $k = 2$	857 (p-value =6.584e-125) df = 91 Verdict: Reject H_N	239 $k = 92$	525 (p-value =1.867e-74) df = 62 Verdict: Reject H_N	425 $k = 63$
IID vs DM	210 (p-value =3.711e-12) df = 87 Verdict: Reject H_N	545 $k = 88$	37 (p-value =0.03568) df = 23 Verdict: Accept H_N	1495 $k = 24$	25 (p-value =0.02672) df = 13 Verdict: Accept H_N	2775 $k = 14$	7 (p-value =0.006668) df = 1 Verdict: Reject H_N	8192 $k = 2$	223 (p-value =3.908e-13) df = 91 Verdict: Reject H_N	62 $k = 92$	171 (p-value =3.807e-12) df = 62 Verdict: Reject H_N	83 $k = 63$
Model vs Measured	9 (p-value =0.009418) df = 2 Verdict: Reject H_N	324 $k = 3$	3 (p-value =0.07644) df = 1 Verdict: Accept H_N	349 $k = 2$	3 (p-value =0.07644) df = 1 Verdict: Accept H_N	349 $k = 2$	12 (p-value =0.03232) df = 5 Verdict: Accept H_N	2701 $k = 6$	107 (p-value =1.876e-13) df = 21 Verdict: Reject H_N	107 $k = 22$	324 (p-value =1.87e-28) df = 89 Verdict: Reject H_N	69 $k = 90$
DM vs Measured	7201 (p-value =0) df = 35 Verdict: Reject H_N	8114 $k = 36$	2554 (p-value =0) df = 12 Verdict: Reject H_N	15349 $k = 13$	543 (p-value =4.795e-114) df = 6 Verdict: Reject H_N	14877 $k = 7$	505 (p-value =8.492e-112) df = 1 Verdict: Reject H_N	2197 $k = 2$	885 (p-value =1.384e-123) df = 105 Verdict: Reject H_N	176 $k = 106$	497 (p-value =1.38e-65) df = 70 Verdict: Reject H_N	331 $k = 71$

TABLE 2. Values obtained for Chi-Squared and MSE analysis for various plots and bin widths using the Substitution Error or No Substitution Error categorisation.

	Low SNR						**High SNR**					
	Bin Width = 1		Bin Width = 5		Bin Width = 10		Bin Width = 1		Bin Width = 5		Bin Width = 10	
	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE
IID vs Measured	4643 (p-value =0) df = 33 Verdict: Reject H_N	5444 $k = 34$	2060 (p-value =0) df = 12 Verdict: Reject H_N	15010 $k = 13$	536 (p-value =1.32e-112) df = 6 Verdict: Reject H_N	14384 $k = 7$	667 (p-value =1.763e-145) df = 2 Verdict: Reject H_N	2916 $k = 3$	673 (p-value =1.821e-91) df = 87 Verdict: Reject H_N	217 $k = 88$	541 (p-value =5.438e-70) df = 79 Verdict: Reject H_N	292 $k = 80$
IID vs DM	162 (p-value =2.049e-06) df = 87 Verdict: Reject H_N	478 $k = 88$	38 (p-value =0.02852) df = 23 Verdict: Accept H_N	2140 $k = 24$	20 (p-value =0.06438) df = 12 Verdict: Accept H_N	1405 $k = 13$	6 (p-value =0.04369) df = 2 Verdict: Accept H_N	656 $k = 3$	148 (p-value =5.249e-05) df = 87 Verdict: Reject H_N	63 $k = 88$	157 (p-value =3.797e-07) df = 79 Verdict: Reject H_N	38 $k = 80$
Model vs Measured	3 (p-value =0.4543) df = 3 Verdict: Accept H_N	132 $k = 4$	2 (p-value =0.1519) df = 1 Verdict: Accept H_N	213 $k = 2$	2 (p-value =0.1519) df = 1 Verdict: Accept H_N	213 $k = 2$	30 (p-value =1.588e-05) df = 5 Verdict: Reject H_N	544 $k = 6$	232 (p-value =4.5e-16) df = 83 Verdict: Reject H_N	52 $k = 84$	258 (p-value =2.053e-18) df = 89 Verdict: Reject H_N	78 $k = 90$
DM vs Measured	4538 (p-value =0) df = 35 Verdict: Reject H_N	5124 $k = 36$	2794 (p-value =0) df = 13 Verdict: Reject H_N	13702 $k = 14$	535 (p-value =2.357e-112) df = 6 Verdict: Reject H_N	14325 $k = 7$	0 (p-value =0.6809) df = 1 Verdict: Accept H_N	181 $k = 2$	725 (p-value =1.13e-111) df = 66 Verdict: Reject H_N	252 $k = 67$	561 (p-value =1.277e-71) df = 84 Verdict: Reject H_N	240 $k = 85$

TABLE 3. Values obtained for Chi-Squared and MSE analysis for various plots and bin widths using the Synchronization Error or No Synchronization Error categorisation.

	Low SNR						**High SNR**					
	Bin Width = 100		Bin Width = 500		Bin Width = 1000		Bin Width = 100		Bin Width = 500		Bin Width = 1000	
	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE
IID vs Measured	239 (p-value =1.331e-52) df = 2 Verdict: Reject H_N	1607 $k = 3$	186 (p-value =1.626e-34) df = 10 Verdict: Reject H_N	408 $k = 11$	187 (p-value =5.082e-37) df = 7 Verdict: Reject H_N	792 $k = 8$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$
IID vs DM	2 (p-value =0.3949) df = 2 Verdict: Accept H_N	7 $k = 3$	8 (p-value =0.5859) df = 10 Verdict: Accept H_N	11 $k = 11$	3 (p-value =0.9043) df = 7 Verdict: Accept H_N	9 $k = 8$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$
Model vs Measured	21 (p-value =0.000893) df = 5 Verdict: Reject H_N	72 $k = 6$	21 (p-value =2.831e-05) df = 2 Verdict: Reject H_N	109 $k = 3$	2 (p-value =0.1328) df = 1 Verdict: Accept H_N	85 $k = 2$	13 (p-value =0.0002457) df = 1 Verdict: Reject H_N	968 $k = 2$	68 (p-value =3.468e-11) df = 9 Verdict: Reject H_N	88 $k = 10$	275 (p-value =2.447e-53) df = 10 Verdict: Reject H_N	559 $k = 11$
DM vs Measured	296 (p-value =5.112e-65) df = 2 Verdict: Reject H_N	1765 $k = 3$	215 (p-value =5.45e-42) df = 8 Verdict: Reject H_N	552 $k = 9$	284 (p-value =6.242e-56) df = 9 Verdict: Reject H_N	686 $k = 10$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$

limitations as only specific, unique errors may be analyzed at a given instance. In this section, a new model is developed which can incorporate memory and multiple types of errors within the channel simultaneously, as opposed to converting the errors into a binary form first.

We again assume the receiver has full knowledge of the transmitted data and that each state emits a unique symbol (**t**, **s**, **d** or **i**) into the synchronization error sequence to identify that particular state ie. that the states are completely visible. This reduces the model to a simple Markov chain. For this analysis, the synchronization error sequence is used directly, where the method of producing the synchronization error sequence is identical to that described in Section III. This sequence, along with the emission matrix, which is a

TABLE 4. Values obtained for Chi-Squared and MSE analysis for various plots and bin widths using the Insertion Error No Insertion Error categorisation.

	Low SNR						High SNR					
	Bin Width = 100		Bin Width = 500		Bin Width = 1000		Bin Width = 100		Bin Width = 500		Bin Width = 1000	
	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE
IID vs Measured	2 (p-value =0.1931) df = 1 Verdict: Accept H_N	72 $k = 2$	2 (p-value =0.1931) df = 1 Verdict: Accept H_N	72 $k = 2$	70 (p-value =1.231e-13) df = 5 Verdict: Reject H_N	109 $k = 6$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$
IID vs DM	0 (p-value =0.7389) df = 1 Verdict: Accept H_N	5 $k = 2$	0 (p-value =0.7389) df = 1 Verdict: Accept H_N	5 $k = 2$	5 (p-value =0.3603) df = 5 Verdict: Accept H_N	7 $k = 6$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$
Model vs Measured	3 (p-value =0.08377) df = 1 Verdict: Accept H_N	93 $k = 2$	3 (p-value =0.2041) df = 2 Verdict: Accept H_N	33 $k = 3$	17 (p-value =0.01051) df = 6 Verdict: Accept H_N	40 $k = 7$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$
DM vs Measured	1 (p-value =0.3374) df = 1 Verdict: Accept H_N	41 $k = 2$	1 (p-value =0.3374) df = 1 Verdict: Accept H_N	41 $k = 2$	39 (p-value =6.034e-07) df = 6 Verdict: Reject H_N	67 $k = 7$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$

TABLE 5. Values obtained for Chi-Squared and MSE analysis for various plots and bin widths using the Deletion Error or No Deletion Error categorisation.

	Low SNR						High SNR					
	Bin Width = 100		Bin Width = 500		Bin Width = 1000		Bin Width = 100		Bin Width = 500		Bin Width = 1000	
	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE	χ^2	MSE
IID vs Measured	4 (p-value =0.05495) df = 1 Verdict: Accept H_N	181 $k = 2$	197 (p-value =2.251e-42) df = 3 Verdict: Reject H_N	354 $k = 4$	102 (p-value =1.588e-20) df = 5 Verdict: Reject H_N	250 $k = 6$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$
IID vs DM	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	8 (p-value =0.04146) df = 3 Verdict: Accept H_N	43 $k = 4$	18 (p-value =0.003366) df = 5 Verdict: Reject H_N	28 $k = 6$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$
Model vs Measured	2 (p-value =0.5285) df = 3 Verdict: Accept H_N	34 $k = 4$	3 (p-value =0.2087) df = 2 Verdict: Accept H_N	42 $k = 3$	11 (p-value =0.02842) df = 4 Verdict: Accept H_N	22 $k = 5$	12 (p-value =0.0004242) df = 1 Verdict: Reject H_N	882 $k = 2$	27 (p-value =5.03e-06) df = 3 Verdict: Reject H_N	124 $k = 4$	151 (p-value =3.068e-29) df = 7 Verdict: Reject H_N	510 $k = 8$
DM vs Measured	4 (p-value =0.04229) df = 1 Verdict: Accept H_N	200 $k = 2$	4 (p-value =0.04229) df = 1 Verdict: Accept H_N	200 $k = 2$	334 (p-value =3.645e-69) df = 6 Verdict: Reject H_N	294 $k = 7$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$	0 (p-value =1) df = 1 Verdict: Accept H_N	0 $k = 2$

diagonal of ones, will produce the transition matrix, A, for a four-state Markov model, which is illustrated in Figure 13, when run through the Baum-Welch algorithm. This process, while simple, produces accurate transition matrices for the memory synchronization channel.

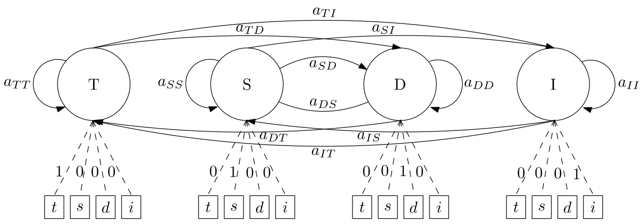


FIGURE 13. Four state Markov model for IDS channel.

A. MEMORY SYNCHRONIZATION CHANNEL MODELS OBTAINED FROM REAL-WORLD DATA

We once again apply this technique to the data from the VLC system [21] with the same parameters as before. Figure 14 shows the transition matrix heatmap for low SNR communication in the VLC system. The actual values for the transition matrix are shown in Equation (6). Using Equations (3) and (5) leads to an error entropy of 0.0247 bits/symbol and consequently a capacity of 0.9753 bits/symbol. From this, a directed graph that represents the Markov chain can be plotted. This is illustrated in Figure 15. It is evident from this model that once a transition occurs, the channel is most likely going to remain in this state. The same thing occurs with substitution errors, as the self-transition probability of this state is quite high. It is also worth noting that the probability

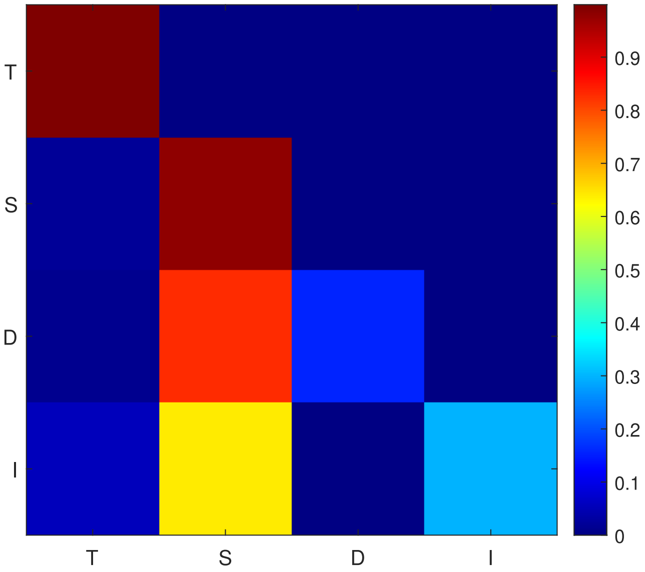


FIGURE 14. Transition matrix heat map at low SNR communication.

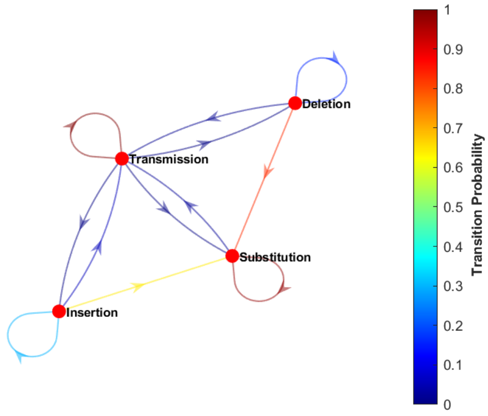


FIGURE 15. Directed graph for proposed IDS memory model at low SNR communication.

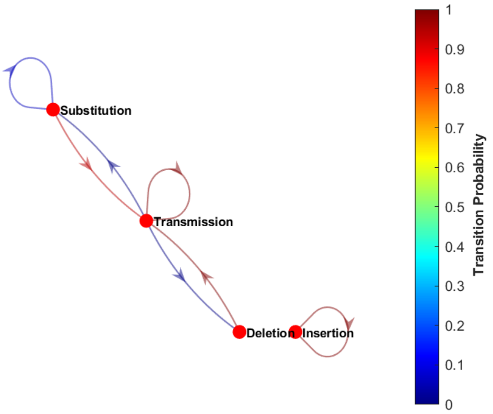


FIGURE 17. Directed graph for proposed IDS memory model at high SNR communication.

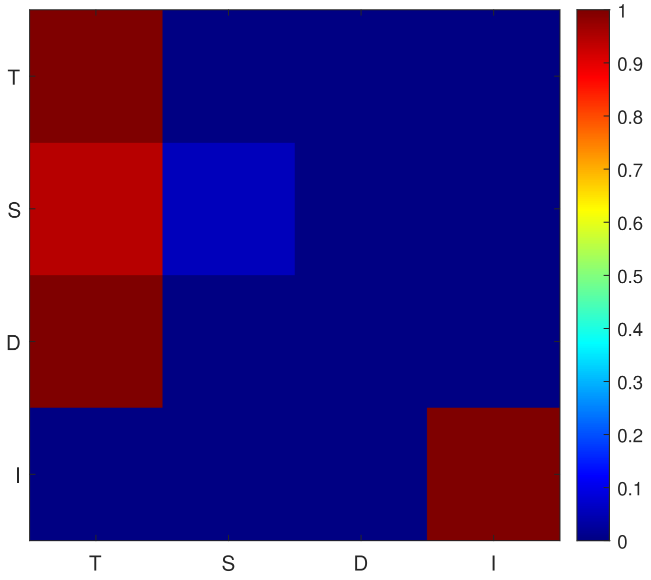


FIGURE 16. Transition matrix heat map at high SNR communication.

of transitioning from a synchronization error (either deletion or insertion) to a substitution error is also quite high, while transitioning from any error back to normal transmission is relatively small. This adequately illustrates the memory we notice between the three error states, as well as the long runs of error-free transmissions observed.

The same analysis is performed for communication at higher SNR values, which yields the transition matrix heatmap shown in Figure 16, and the directed graph shown in Figure 17. Equation (7) represents the transition matrix obtained for high SNR communication. Equation (3) unfortunately cannot be used to determine the entropy in this case because the Markov chain is not ergodic as it is reducible and contains states that are isolated and absorbing in nature. It is, however, still evident from the high SNR communication analysis that once the system is in an error-free (transmission) state, it will more than likely continue in this state. Even if a substitution error occurs, the system quickly returns to the transmission state with very little probability of consecutive errors. It is even shown that as soon a single deletion occurs, the system still immediately returns to error-free transmission. It is worth noting that since no insertions are observed at high SNR, we assume there are no transitions out of this state, hence the insertion state is isolated in Figure 17. Once again the analysis of the real-world data corroborates our notion of when the memory models may be beneficially used. As expected, from the high SNR transition matrix shown in Equation (7), it is noticeable that the channel degenerates into a memoryless IID model where the transitions are likely to stay in the error-free state while sporadically transitioning to a substitution-error state before quickly returning to an error-free transmission. This additionally provides a useful method in determining if a system inherently contains memory as this captured memory is clearly indicated by the values obtained in the respective transition matrix.

$$A_{LowSNR} = \begin{bmatrix} 0.9986 & 0.0011 & 1.535 \times 10^{-4} & 1.089 \times 10^{-4} \\ 0.0203 & 0.9797 & 0 & 0 \\ 0.0122 & 0.8293 & 0.1585 & 0 \\ 0.0571 & 0.6429 & 0 & 0.3 \end{bmatrix} \quad (6)$$

$$A_{HighSNR} = \begin{bmatrix} 0.9980 & 0.0020 & 1.0120 \times 10^{-6} & 0 \\ 0.9421 & 0.0579 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (7)$$

VI. CONCLUSION

Various memory models and synchronization error channels are discussed, but there is, unfortunately, no overlap which accounts for IDS channels that contain statistically dependent errors. Firstly, a novel technique to determine channel characteristics and model parameters is introduced which builds onto the idea of the Fritchman model while making use of the Levenshtein distance and different error categories. The proposed channel models show a clear distinction for low SNR communication in the error and error-free runs from the DM channel, which seeks to only model statistically independent synchronization errors. Finally, a more encompassing model, which can be viewed as a Markov chain, that accounts for insertion, deletion, and substitution errors is described. The new method is more enveloping of practical, real-world communication channels and is demonstrated by making use of data from a VLC system. This method additionally serves as a way to indicate if a channel inherently contains correlated errors (synchronization or otherwise) and it is evident that at high SNR communication, the model degenerates into an IID channel. This technique may even be applied to other channels in applications such as the barcoding of DNA sequences.

APPENDIX.

A. PERFORMANCE METRICS FOR ANALYSIS AND METHODOLOGY

To compare and quantify the similarities, and consequently, differences, between the various plots produced; the Chi-Square (χ^2) and Mean Squared Error (MSE) metrics are used. This Appendix outlines the background on the various analysis metrics used as well as the procedures and methods used when calculating these quantities.

1) CHI-SQUARED (χ^2) GOODNESS OF FIT

The Chi-Squared test is a non-parametric test that is used to determine if the observed data is significantly different from expected data and is calculated using Equation (A.1) where k is the number of data samples, o_i is the observed data at sample i and e_i is the expected value of the ith sample [26]–[28].

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \tag{A.1}$$

As this is a method used in hypothesis testing, two hypotheses first need to be constructed and later tested. The null hypothesis H_N describes the situation where there is no significant difference in the distributions of the compared plots whereas the alternative hypothesis, H_A, describes the situation where there is a significant difference in distribution between the compared plots. For all the tests, a strict significance value of 0.01 is used, but it is not uncommon to use 5% or 10% depending on the scenario. The degrees of freedom vary as a function of the bin size and the context of the plots.

2) MEAN SQUARED ERROR (MSE)

The Mean Squared Error is a metric that is traditionally used to indicate how close a regression line is to a set of observed values [29]. Here we will use the MSE, in a similar fashion to the χ^2 metric, as an indicator of how similar the various error-free run plots are. The MSE equation is described by Equation (A.2) where the variables are defined as before for the χ^2 equation. Unlike the Chi-Squared metric, there is no way of determining a good fit with the MSE value alone and thus we will need to compare different MSE values from different plots against each other, where a larger value of the MSE indicates a more significant difference between the plots compared.

$$MSE = \frac{1}{k}\sum_{i=1}^{k}(o_i - e_i)^2 \tag{A.2}$$

3) ERROR-FREE RUN AND ERROR RUN DISTRIBUTIONS

Two of the most important metrics used to define discrete channel models are the error-free run distribution and the error run distribution. The error-free run distribution is defined as $Pr(0^m|1)$, which is the probability of receiving a stream of m or more consecutive error-free transmissions following an error. An error run distribution is defined as $Pr(1^m|0)$ and describes the probability of receiving m or more consecutive errors after an error-free transmission [17].

In general, for the Fritchman model, the error-free runs and error runs are given by Equations (A.3) and (A.4) respectively, which describes these distribution events in terms of weighted exponentials [11], [17]. Here, λ_i represents the eigenvalues of A_{gg} and Abb where A_{gg} and Abb are the diagonal sub-matrices which form part of the transition matrix A and f_i is the corresponding transition probability from a_{ij} [11], [17]. For a single error state Fritchman model, Equation (A.3) is simplified to Equation (A.5) [17].

$$Pr(0^m|1) = \sum_{i=1}^{k} f_i\lambda_i^{m-1} \tag{A.3}$$

$$Pr(1^m|0) = \sum_{i=k+1}^{k} f_i\lambda_i^{m-1} \tag{A.4}$$

$$Pr(0^m|1) = \sum_{k=1}^{N-1} \frac{a_{Nk}(a_{kk})^m}{a_{kk}} \tag{A.5}$$

4) PROCEDURE FOR χ^2 AND MSE CALCULATIONS

For both the χ^2 and the MSE, the data from the error-free runs is segmented into different bins according to a predefined bin width. Bin widths of 1, 5 and 10 are chosen for Error vs Error-Free segmentation as well as Substitution Error vs No Substitution Error segmentation while bin widths of 100, 500 and 1000 are selected for the remaining segmentation plots. The rationale behind this is to ensure that the large counts condition is met. The large counts condition states that each category (in our case each bin) has an expected outcome of at least 5 [26]. This is done to ensure the criteria of the χ^2 analysis is met as five or more occurrences in each expected bin satisfies the criteria for the central limit theorem which allows the distributions to be normally distributed in nature ie. there needs to be a large enough samples for the central limit theorem to be met as the χ^2 statistic is based on a normal distribution [27]. Additionally, the traditional definition of the error-free runs, $\Pr(0^m|1)$, which is the probability of receiving a stream of **at least** m consecutive error-free transmissions following an error is replaced by a stricter criteria in the χ^2 and the MSE analysis where the bins are categorised by **exactly** m error-free transmissions following an error. In other words, if the bin width is set to 5, we would have categories or bins of 1 to 5 error-free transmissions, 6 to 10 error-free transmissions, 11 to 15 error-free transmissions and so on. Here the bin 1 to 5 error-free transmissions does not represent at least 1 to 5 error-free transmissions but rather how many times exactly 1,2,3,4 or 5 error-free transmissions occur. This is done to satisfy the criteria that $\sum_{i=1}^{k} r_i = 1$ where r_i is the proportion or percentage of counts in each category [28]. In other words, all counts are independent and cannot attribute to multiple categories. Finally, a cutoff for the number of consecutive error-free runs must be set as many of the longer chains of error-free transmissions become less frequent and often times the counts are less than five occurrences. Again, the rationale is to satisfy the large counts criteria so that the χ^2 analysis is valid. As the longer runs of error-free transmissions are less likely to occur in the expected variable, the first bin that has a frequency or occurrence count less than 5 is selected as the cutoff value and all subsequent categories, including the cutoff bin, are grouped together as a single bin. Once the segmentation of data into bins is complete the χ^2 and MSE statistics may now be calculated using Equation (A.1) and Equation (A.2) respectively.

For the χ^2 statistic, the p-value is then calculated using the degrees of freedom which corresponds to the number of bins less one. Once the p-value is obtained we can then compare this to the significance level originally chosen for the analysis. If the p-value is less than the significance level, the null hypothesis is rejected and consequently, the alternate hypothesis is accepted [26]–[28]. If the p-value is greater than the significance level, it implies there is not enough evidence to suggest the null hypothesis is wrong and thus we accept it as true. It is worth noting that as the bin width increases, the degrees of freedom (categories) decreases. In some cases, there is only one bin that contains all counts and the corresponding Chi-Square and MSE values tell us if the total expected number of errors is in agreement with the total measured number of errors for the given parameters. Furthermore, we note that more accurate values for these metrics will be obtained when using optimised bin widths for given parameters.

REFERENCES

[1] D. Kracht and S. Schober, "Using the Davey–MacKay code construction for barcodes in DNA sequencing," in *Proc. 8th Int. Symp. Turbo Codes Iterative Inf. Process. (ISTC)*, Aug. 2014, pp. 142–146.

[2] D. Kracht and S. Schober, "Insertion and deletion correcting DNA barcodes based on watermarks," *BMC Bioinf.*, vol. 16, no. 1, pp. 1–14, Dec. 2015.

[3] M. C. Davey and D. J. C. Mackay, "Reliable communication over channels with insertions, deletions, and substitutions," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 687–698, Feb. 2001.

[4] R. G. Gallager, "Sequential decoding for binary channels with noise and synchronization errors," Massachusetts Inst. Technol., Lincoln Lab, Lexington, MA, USA, Tech. Rep. 2502, 1961.

[5] K. Zigangirov, "Sequential decoding for a binary channel with drop-outs and insertions," *Problemy Peredachi Inf.*, vol. 5, no. 2, pp. 23–30, 1969.

[6] M. C. Davey, "Error-correction using low-density parity-check codes," Ph.D. dissertation, Inference Group, Cavendish Lab., Univ. Cambridge, Cambridge, U.K., 2000.

[7] L. N. Kanal and A. R. K. Sastry, "Models for channels with memory and their applications to error control," *Proc. IEEE*, vol. 66, no. 7, pp. 724–744, Jul. 1978.

[8] P. Sadeghi, R. Kennedy, P. Rapajic, and R. Shams, "Finite-state Markov modeling of fading channels—A survey of principles and applications," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 57–80, Sep. 2008.

[9] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Syst. Tech. J.*, vol. 39, no. 5, pp. 1253–1265, Sep. 1960.

[10] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J.*, vol. 42, no. 5, pp. 1977–1997, Sep. 1963.

[11] B. Fritchman, "A binary channel characterization using partitioned Markov chains," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 2, pp. 221–227, Apr. 1967.

[12] D. Leigh. (2001). *Capacity of Insertion and Deletion Channels, Project Report*. [Online]. Available: http://www.inference.phy.cam.ac.uk/is/papers

[13] F. Wang, "Coding for insertion/deletion channels," Ph.D. dissertation, School Elect., Comput. Energy Eng., Arizona State Univ., Tempe, AZ, USA, 2012.

[14] F. Swarts, "Markov characterization of fading channels," Ph.D. dissertation, Cybern. Lab., Univ. Johannesburg, Johannesburg, South Africa, 1991.

[15] J. G. Proakis, *Digital Communications*. New York, NY, USA: McGraw-Hill, 2001, pp. 777–778.

[16] L. Zhong, F. Alajaji, and G. Takahara, "A binary communication channel with memory based on a finite queue," *IEEE Trans. Inf. Theory*, vol. 53, no. 8, pp. 2815–2840, Aug. 2007.

[17] W. H. Tranter, T. S. Rappaport, K. L. Kosbar, and K. S. Shanmugan, *Principles of Communication Systems Simulation With Wireless Applications*, vol. 1. Upper Saddle River, NJ, USA: Prentice-Hall, 2004.

[18] D. G. Holmes, L. Cheng, M. Shimaponda-Nawa, A. D. Familua, and A. M. Abu-Mahfouz, "Modelling noise and pulse width modulation interference in indoor visible light communication channels," *AEU, Int. J. Electron. Commun.*, vol. 106, pp. 40–47, Jul. 2019.

[19] K. Shanmugam, *Digital and Analog Communication Systems*. New York, NY, USA: Wiley, 1979.

[20] S. Achari. *TTL-VLC Data*. Accessed: Jan. 18, 2021. [Online]. Available: https://github.com/WitsOCLab/TTL-VLC

[21] S. Achari, A. Yi Yang, J. Goodhead, B. Swanepoel, and L. Cheng, "Self-synchronising on-off-keying visible light communication system for intra and inter-vehicle data transmission," 2021, *arXiv:2101.05126*. [Online]. Available: http://arxiv.org/abs/2101.05126

[22] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys.-Dokl.*, vol. 10, no. 8, pp. 707–710, 1966.

[23] M. Gilleland. *Levenshtein Distance, in Three Flavors*. Accessed: Feb. 19, 2021. [Online]. Available: https://people.cs.pitt.edu/kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein%20Distance.htm

[24] GeeksforGeeks. *Dynamic Programming | Set 5 (Edit Distance)*. Accessed: Feb. 19, 2021. [Online]. Available: http://www.geeksforgeeks.org/dynamic-programming-set-5-edit-distance

[25] D. G. Holmes, "Semi-hidden Markov models for visible light communication channels," M.S. thesis, School Elect. Inf. Eng., Univ. Witwatersrand, Johannesburg, Johannesburg, South Africa, 2018.

[26] Khan Academy. *Chi-Square Goodness-of-Fit Example*. Accessed: Nov. 25, 2020. [Online]. Available: https://www.khanacademy.org/math/ap-statistics/chi-square-tests/chi-square-goodness-fit/v/goodness-of-fit-example

[27] J. Zeltzer. *Zedstatistics—Chi-Squared Goodness of Fit Test! Extensive Video!* Accessed: Nov. 25, 2020. [Online]. Available: https://youtu.be/ZNXso_riZag

[28] G. Paul, *Introductory Statistics for the Social Sciences*. Regina, SK, Canada: Univ. of Regina, Department of Sociology and Social Sciences, 1992.

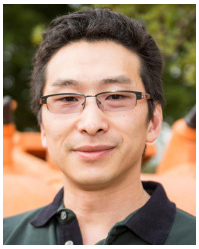
[29] M. Binieli. *Machine Learning: An Introduction to Mean Squared Error and Regression Lines*. Accessed: Nov. 25, 2020. [Online]. Available: https://www.freecodecamp.org/news/machine-learning-mean-squared-error-regression-line-c7dde9a26b93

SHAMIN ACHARI received the B.Sc. degree in electrical and information engineering from the University of the Witwatersrand (WITS), Johannesburg, South Africa, in 2015, where he is currently pursuing the Ph.D. degree in electrical engineering.

His M.Sc. was based in the field of visible light communication with special focus on error correction schemes for visible light systems, which was subsequently upgraded to the Ph.D. with a focus on channel modeling and methods of correcting synchronization errors. His research interests include visible light communications, machine learning, and the Internet of Things (IoT).

DANIEL G. HOLMES received the B.Sc.Eng. degree (Hons.) in electrical and information engineering and the M.Sc.Eng. degree in visible light communication channel modeling from the University of the Witwatersrand, Johannesburg, in 2015 and 2018, respectively.

He was awarded the Adolf Goldsmith Memorial Fund Prize by the Wits School of Electrical and Information Engineering, in 2015, in recognition of the significant progress made in his undergraduate degree. He served as the Chair for the SAIEE-IEEE Wits Student Branch, from 2016 to 2017.



LING CHENG (Senior Member, IEEE) received the B.Eng. degree *(cum laude)* in electronics and information from the Huazhong University of Science and Technology (HUST), in 1995, the M.Ing. degree *(cum laude)* in electrical and electronics, in 2005, and the D.Ing. degree in electrical and electronics from the University of Johannesburg (UJ), in 2011.

In 2010, he joined the University of the Witwatersrand, where he was promoted to Full Professor, in 2019. He has been a visiting professor with five universities and the principal advisor for over 40 full research master's students. He has published more than 100 research articles in journals and conference proceedings. His research interests include telecommunications and artificial intelligence. He was awarded the Chancellor's Medals in 2005 and 2019, the National Research Foundation rating, in 2014, and the Best Student Paper Award from the IEEE ISPLC, Austin, in 2015, which was made to his Ph.D. student. He is the Vice-Chair of the IEEE South African Information Theory Chapter. He serves as an Associate Editor for three journals.