Masked Symbol Modeling for Demodulation of Oversampled Baseband Communication Signals in Impulsive Noise-Dominated Channels

Oguz Bedir*
Electrical & Computer Engineering
Texas A&M University
College Station, TX 77843
oguzbedir@tamu.edu

Nurullah Sevim*
Electrical & Computer Engineering
Texas A&M University
College Station, TX 77843
nurullahsevim@tamu.edu

Mostafa Ibrahim
Engineering Technology &
Industrial Distribution
Texas A&M University
College Station, TX 77843
mostafa.ibrahim@tamu.edu

Sabit Ekin
Engineering Technology &
Industrial Distribution, and
Electrical & Computer Engineering
Texas A&M University
College Station, TX 77843
sabitekin@tamu.edu

Abstract

Recent breakthroughs in natural language processing show that attention mechanism in Transformer networks, trained via masked-token prediction, enables models to capture the semantic context of the tokens and internalize the grammar of language. While the application of Transformers to communication systems is a burgeoning field, the notion of context within physical waveforms remains under-explored. This paper addresses that gap by re-examining inter-symbol contribution (ISC) caused by pulse-shaping overlap. Rather than treating ISC as a nuisance, we view it as a deterministic source of contextual information embedded in oversampled complex baseband signals. We propose Masked Symbol Modeling (MSM), a framework for the physical (PHY) layer inspired by Bidirectional Encoder Representations from Transformers methodology. In MSM, a subset of symbol-aligned samples is randomly masked, and a Transformer predicts the missing symbol identifiers using the surrounding "in-between" samples. Through this objective, the model learns the latent syntax of complex baseband waveforms. We illustrate MSM's potential by applying it to the task of demodulating signals corrupted by impulsive noise, where the model infers corrupted segments by leveraging the learned context. Our results suggest a path toward receivers that interpret, rather than merely detect communication signals, opening new avenues for context-aware PHY layer design.

1 Introduction

The success of large language models (LLMs) in natural language processing (NLP) is largely attributed to the Transformer architecture [1]. Its attention mechanism enables models to capture long-range dependencies and understand the semantic context in which words appear. This paradigm

*These authors contributed equally to this work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

has fueled successful applications in diverse fields [2–4], including communication systems [5–8]. However, while Transformers are increasingly used to solve complicated communication problems, the fundamental notion of context within physical waveforms remains largely unexplored. This paper aims to bridge that gap.

The overlap between adjacent pulses, inherent in pulse-shaping, embeds a predictable structure in the oversampled baseband signal. Each sample, therefore, contains information not only about its primary symbol but also about its neighbors, creating what we term inter-symbol contributions (ISCs). In oversampled systems, ISCs have typically been harnessed within equalization methods, but their broader potential remains under-explored. We reframe them instead as a deterministic source of information that a sophisticated model can exploit to learn contextual representations.

To exploit this structure, we introduce Masked Symbol Modeling (MSM), a self-supervised, Bidirectional Encoder Representations from Transformers (BERT)-style [9] framework designed for the physical (PHY) layer. MSM trains a Transformer to predict randomly masked symbols by analyzing the surrounding unmasked waveform samples. Through this objective, the model learns to internalize the underlying structure of pulse-shaped signals, moving beyond simple detection towards a more comprehensive interpretation of the waveform.

Our main contributions are as follows:

- We introduce MSM, a novel framework for learning representations of oversampled complex-valued baseband signals by treating ISC as a source of contextual information.
- We demonstrate the practical utility of MSM by applying it to symbol prediction under impulsive noise (e.g., Middleton Class-A [10, 11]), where the model leverages learned context to recover corrupted symbols.

Our results suggest a path toward receivers that interpret, rather than merely detect, communication signals, opening new avenues for designing intelligent and context-aware PHY layer.

The rest of the paper is organized as follows: Section 2 introduces the signal and noise model used in this work. Section 3 presents the proposed MSM framework and its architectural details. Section 4 describes the experimental setup and presents our results. Section 5 concludes the paper and discusses future directions.

2 Preliminaries

2.1 Digital Communication Basics

Let $\{x[k]\} \in \mathbb{C}^N$ be a symbol sequence, drawn from a modulation set $\mathcal{M}$, and $g(t)$ the pulse shaping filter. The continuous-time baseband signal can be expressed as [12]

$$s(t) = \sum_{k=0}^{N-1} x[k] \cdot g(t - kT), \tag{1}$$

where T is the symbol duration. Sampling at L samples per symbol (SPS) produces the corresponding discrete-time baseband signal

$$s[n] = s(nT_s) = \sum_{k=0}^{N-1} x[k] \cdot g(nT_s - kT), \tag{2}$$

where $T_s = \frac{T}{L}$.

2.2 Noise Model

A widely used impulsive-noise model is Middleton Class-A. The probability density function (PDF) of a real-valued noise sample n_i is an infinite mixture of Gaussian distributions with the mixture weights following a Poisson distribution [13–15]:

$$f(n_i) = \sum_{m=0}^{\infty} \frac{A^m \exp(-A)}{m!} \frac{1}{\sigma_m \sqrt{2\pi}} \exp\Big(- \frac{n_i^2}{2\sigma_m^2} \Big), \tag{3}$$

where $m \sim \text{Poisson}(A)$, $n_i|m \sim \mathcal{N}(0, \sigma_m^2)$ with $\sigma_m^2 = \sigma_g^2 \left(\frac{m}{A\Gamma} + 1\right), \Gamma = \frac{\sigma_g^2}{\sigma_I^2}$. Here σ_g^2 is the background Gaussian noise power, σ_I^2 is the impulsive-noise power, and the mean total power is $\sigma_{\text{total}}^2 = \sigma_g^2 + \sigma_I^2$. For complex baseband, apply the real-valued model independently to I/Q, yielding a circularly symmetric complex process.

3 Methodology

This work investigates how the deterministic structure introduced by pulse shaping can be exploited to learn meaningful representations of communication signals. In particular, we hypothesize that the ISC arising from pulse shaping introduces a form of contextual dependency between samples that can be effectively modeled using attention-based architectures. Fig. 1 illustrates how overlapping symbol contributions create such structured sample-level context.

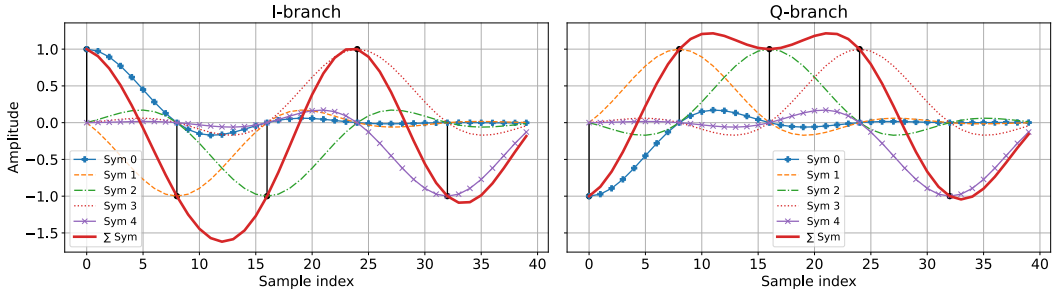


Figure 1: Visualization of ISC introduced by pulse shaping. Each colored segment represents the contribution of a distinct symbol to the overall waveform. Due to pulse overlap, each sample contains information from multiple adjacent symbols, creating structured context that can be exploited.

We adopt a BERT-style training framework with a Transformer neural network (NN). The model input is a sequence of complex-valued baseband symbols, which are mapped to oversampled time-domain waveforms using pulse shaping filters. During training, 15% of the symbols are randomly masked by setting their corresponding sample spans to zero. A discrete vocabulary assigns a unique symbol identifier (ID) to each distinct constellation point across all considered modulations. The model is trained to predict the correct ID for each masked symbol using only surrounding unmasked samples, thereby internalizing the contextual structure imposed by pulse shaping and ISC (Fig. 2).

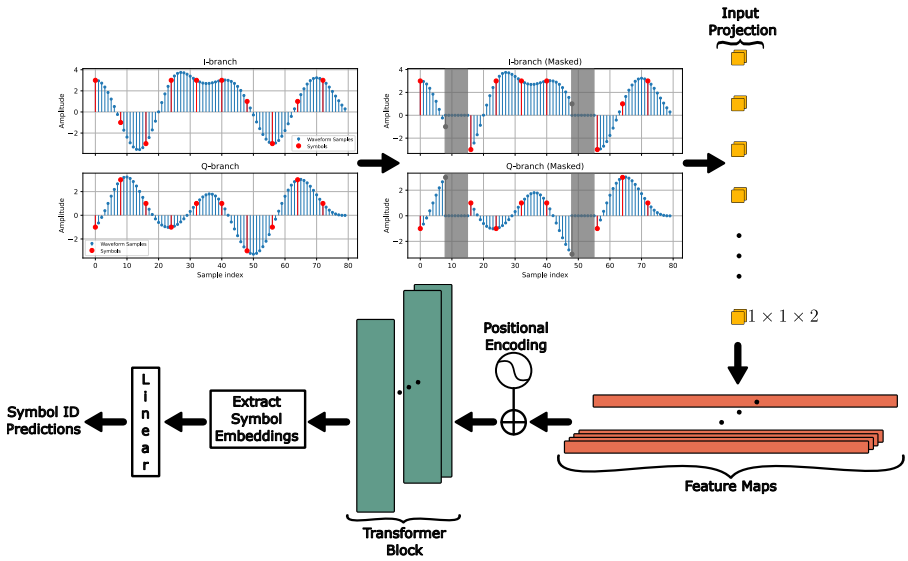


Figure 2: Conceptual diagram of end-to-end MSM architecture.

While the primary objective is to study the contextual properties of pulse-shaped signals, we additionally demonstrate the utility of the model by applying it to recover symbols corrupted by impulsive noise.

3.1 Training and Inference Signals

For training, the model uses the non-impaired basebands signal $s[n]$ as defined in Sec. 3.1. At inference, we evaluate the model on signals corrupted by additive Middleton Class-A noise, $z[n]$, resulting in the received signal:

$$y[n] = s[n] + z[n], \quad z[n] \sim \text{Middleton Class-A}(A, \Gamma). \tag{4}$$

Table 1: Dataset description.

Description	**Range**
Modulation types	{BPSK, QPSK, PSK8, PSK16, QAM4, QAM16, QAM64, QAM256}
Symbol rate (symbols/s)	1
SPS	8
Filter span (symbols)	{10, 12, 14, 16}
Roll-off factor	{0.25, 0.35, 0.45, 0.55, 0.65, 0.75}

3.2 Data Generation Pipeline

Waveforms are generated on-the-fly using a modular NumPy/PyTorch pipeline. For each example, a modulation scheme is sampled, symbols are drawn uniformly from the corresponding constellation, and mapped to I/Q components. These symbols are represented as two real-valued arrays (I and Q) rather than a single complex array. Each branch is pulse-shaped with a raised cosine (RC) finite impulse response (FIR) filter, with span and roll-off uniformly sampled from Table 1. Waveforms are normalized to unit power, with SPS fixed at 8 to ensure 1024-sample inputs for the Transformer. Each waveform is paired with a target sequence corresponding to correct symbol IDs for supervised training. These symbol sequences serve as the ground-truth labels for masked positions during training. Training uses only non-impaired waveforms. At inference, impulsive noise can be optionally added using a Middleton Class-A noise model in a fully vectorized implementation. Code is available at the following repository: `https://github.com/OguzBedir/Masked_Symbol_Modeling.git`.

3.3 System Model

Each training waveform passes through a masking module that zeros the samples of a random 15% of symbols. The masked signal is processed by the `Masked Symbol Transformer`, which maps the two input channels to a $512-$dimensional embedding via a learnable 1D projection layer, adds fixed sinusoidal positional encoding, and applies six Reformer blocks with locality-sensitive hashing (LSH) attention (bucket size 64, four hashes), shared weights, and reversible layers for memory efficiency. For each masked symbol, embeddings over its sample span are mean-pooled and passed to a linear classifier mapping $\mathbb{R}^{512} \to \mathbb{R}^{V}$, where $V = 272$ is the vocabulary size. Loss is computed with cross-entropy only over masked symbols, with inverse-frequency weighting to mitigate class imbalance.

3.4 Training Setup

Training is fully self-supervised, using an on-the-fly `IterableDataset`. No external datasets or pre-recorded signals are used. Optimization uses Adam (10^{-3} learning rate) on all parameters jointly. Experiments are run on a single NVIDIA A100 for 24 hours, corresponding to $37{,}551$ training steps, with a batch size of 64. As data are generated procedurally, there is no notion of epochs.

4 Experimental Setup & Results

We evaluate MSM for predicting symbols affected by impulsive noise, demonstrating its utility in a representative application. All experiments use the same base waveform configuration as in training: each waveform has 8 SPS, a filter span (in symbols) and a roll-off factor uniformly sampled from the corresponding values listed in Table 1. Each waveform consists of 1024 samples, and is trimmed to remove leading and trailing transients introduced by filter delays, ensuring alignment between pulse-shaped waveforms and symbol boundaries, identical to the training configuration. The embedding dimension is fixed at 512. To ensure statistical reliability and avoid favorable outcomes due to random initialization, all experiments in this section are repeated with three different random seeds, and the average performance is reported. For each seed, the simulation is run on at least $10{,}000$ target symbols.

We first evaluate MSM on non-impaired signals, under conditions identical to those used in training: 15% of the symbols per waveform are randomly masked, and the model predicts the corresponding symbol IDs. Results are reported separately for each modulation scheme, as well as for a "mixed" setting in which the modulation scheme is sampled uniformly at random from Table 1. Each waveform contains 128 symbols, and inference is performed in batches of 64 waveforms. With 15% masking, this yields 19 masked symbols per waveform, or $1{,}216$ prediction targets per batch. Nine batches are processed per seed, resulting in $10{,}944$ predictions. As expected, the model achieves higher accuracy on simpler modulation formats, with performance remaining stable across most schemes as shown in Fig. 3. Notable deviations occur for Binary Phase-Shift Keying (BPSK) and Quadrature Amplitude Modulation (QAM)-256, where accuracy is significantly higher or lower, respectively, due to the extreme simplicity or complexity of their constellations.

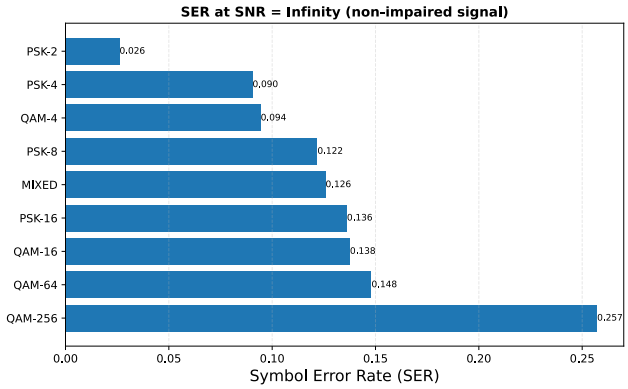
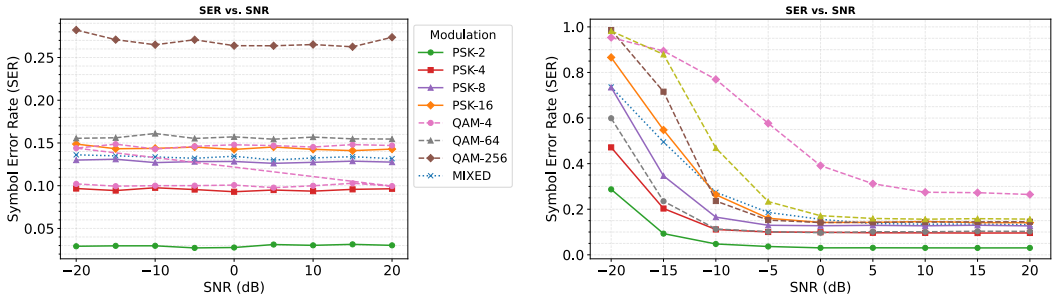


Figure 3: Symbol Error Rate (SER) for different modulation schemes at infinite Signal-to-Noise Ratio (SNR) (i.e., non-impaired signals).

To assess utility under impairments, we evaluate the model on waveforms impaired by Middleton Class-A impulsive noise. The fully corrupted waveform is not fed directly to the model; instead, symbols affected by impulsive noise are identified, and their corresponding waveform segments are masked, as illustrated in Fig. 2, while unaffected segments are left unchanged. The model is then tasked with recovering the symbol IDs at these masked positions using contextual information from surrounding unmasked samples.

This evaluation represents a semi-synthetic impairment scenario, designed to isolate the model's ability to exploit contextual structure rather than to assess robustness against widespread noise. We use $\Gamma \in \{10^{-3}, 10^{-6}\}$ [14] to ensure that the Gaussian background noise component is insignificant and that the impairment is dominated by impulsive noise. For both cases in Fig. 4, the impulsive index is set to $A = \frac{-\ln(0.85)}{L}\big|_{L=8}$ corresponding to an average symbol-hit rate of 15%.

Since the noise generation process is modeled faithfully according to the Middleton Class-A distribution, it remains inherently stochastic. Consequently, achieving exactly 15% symbol hits is not possible unless (i) high-cost rejection sampling is applied, or (ii) a direct symbol-selection approach is used, which would violate proper statistical noise modeling. The derivation of the chosen A value and the computation of its confidence interval are provided in Appendix A.1.



(a) SER under strong impulsive noise ($\Gamma = 10^{-6}$). (b) SER under moderate impulsive noise ($\Gamma = 10^{-3}$).

Figure 4: Symbol Error Rate (SER) of the proposed model under Middleton Class-A noise for two Γ values with A set for an average 15% symbol-hit rate.

In the configuration shown in Fig. 4a, the Gaussian-to-impulsive noise ratio Γ is so small that the Gaussian component is effectively negligible; and impulsive bursts dominate the Signal-to-Noise Ratio (SNR). Because the masking removes affected symbols and Gaussian noise power is minimal, performance remains nearly constant across SNR values.

The behavior differs in Fig. 4b, where Γ is larger and the Gaussian component is no longer negligible. In the low-SNR regime, Gaussian noise significantly degrades performance, even after masking impulsive noise–affected symbols. As the SNR increases, the influence of the Gaussian component diminishes, and performance stabilizes; similar to that observed in the first configuration.

5 Conclusion

This work investigates how the deterministic contextual structure introduced by pulse shaping can be exploited for symbol inference. We demonstrate that Transformer networks, through their attention mechanism, can leverage this structured context to predict symbol IDs from surrounding "in-between" samples. The results support our central hypothesis that it is possible to design receivers that go beyond conventional symbol detection and instead interpret the transmitted waveform structure to recover information. This capability may enhance error correction and other downstream tasks. While promising, the findings are preliminary and require systematic ablation to fully understand trade-offs and limitations.

A key direction for improvement is refining the input representation. In our current setup, the model processes signals in their native $\mathbb{R}^2$ form, with separate I and Q channels. An alternative is to quantize the amplitudes and define the vocabulary over quantization-level pairs (I, Q), assigning a distinct embedding vector to each pair, including a dedicated "mask" embedding. All embeddings, including the mask, would be initialized randomly and learned during training, enabling the model to develop richer, semantically meaningful representations. When trained across diverse signal and channel conditions, these embeddings could encode both intrinsic waveform properties and channel characteristics. Such changes would make the framework more faithful to the original BERT formulation and could remove the need for the current input projection layer.

Learning the mask embedding would encourage a more expressive latent space, enabling operation under more challenging conditions. Explicit masking at inference, used in our current setup, may then be unnecessary, as the model could process unmasked waveforms directly while retaining performance. Another direction is shifting from symbol-level to sample-level prediction, potentially improving contextual modeling and robustness.

Future ablations should examine: (i) embedding dimension, (ii) Transformer depth, attention heads, and encoder/decoder variants, and (iii) masking strategies, including percentage and learned versus fixed mask embeddings, and (iv) a comparison with alternative deep learning architectures that do not use an attention mechanism to determine whether attention provides substantial gains in leveraging signal context. These studies will clarify how representation, architecture, and masking interact, guiding broader applicability to diverse and challenging channels.

Acknowledgments and Disclosure of Funding

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Early Career Research Program under Award Number DE-SC0023957.

Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

[3] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. Nelson, Alex Bridgland, and et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, Jan 2020. doi: 10.1038/s41586-019-1923-7.

[4] Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778, 2018. doi: 10.1109/ICASSP.2018.8462105.

[5] Yoni Choukroun and Lior Wolf. Error correction code transformer. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 38695–38705. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/fcd3909db30887ce1da519c4468db668-Paper-Conference.pdf`.

[6] Zhuolin Chen, Fanglin Gu, and Rui Jiang. Channel estimation method based on transformer in high dynamic environment. In *2020 International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 817–822, 2020. doi: 10.1109/WCSP49889.2020.9299821.

[7] Matteo Zecchin, Kai Yu, and Osvaldo Simeone. In-context learning for mimo equalization using transformer-based sequence models. In *2024 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1573–1578, 2024. doi: 10.1109/ICCWorkshops59551.2024.10615360.

[8] Jonathan Ott, Jonas Pirkl, Maximilian Stahlke, Tobias Feigl, and Christopher Mutschler. Radio foundation models: Pre-training transformers for 5g-based indoor localization. In *2024 14th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–6, 2024. doi: 10.1109/IPIN62893.2024.10786154.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423/`.

[10] David Middleton. Statistical-physical models of electromagnetic interference. *IEEE Transactions on Electromagnetic Compatibility*, EMC-19(3):106–127, 1977. doi: 10.1109/TEMC.1977.303527.

[11] Leslie A. Berry. Understanding middleton's canonical formula for class a noise. *IEEE Transactions on Electromagnetic Compatibility*, EMC-23(4):337–344, 1981. doi: 10.1109/TEMC.1981.303965.

[12] Emil Björnson and Özlem Tuğfe Demir. *Introduction to Multiple Antenna Communications and Reconfigurable Surfaces*. Now Publishers, Boston – Delft, 2024. ISBN 978-1-63828-314-0. doi: 10.1561/9781638283157.

[13] Hyungkook Oh and Haewoon Nam. Simple calculation of thresholds for adaptive modulation in middleton class a noise. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2017. doi: 10.1109/WCNC.2017.7925923.

[14] Jing Lin, Marcel Nassar, and Brian L. Evans. Impulsive noise mitigation in powerline communications using sparse bayesian learning. *IEEE Journal on Selected Areas in Communications*, 31(7):1172–1183, 2013. doi: 10.1109/JSAC.2013.130702.

[15] Thokozani Shongwey, A. J. Han Vinck, and Hendrik C. Ferreira. On impulse noise and its models. In *18th IEEE International Symposium on Power Line Communications and Its Applications*, pages 12–17, 2014. doi: 10.1109/ISPLC.2014.6812360.

[16] Russell Impagliazzo and Valentine Kabanets. Constructive proofs of concentration bounds. In Maria Serna, Ronen Shaltiel, Klaus Jansen, and José Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 617–631, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15369-3.

[17] Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. *Concentration Inequalities*, pages 208–240. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9. doi: 10.1007/978-3-540-28650-9_9. URL `https://doi.org/10.1007/978-3-540-28650-9_9`.

A Impulsive Index: Calibration and Concentration

A.1 Calibrating the Impulsive Index

Let $\{m[n]\}_{n=0}^{N-1}$ denote the per-sample impulsive counts in the Middleton Class-A model, with $m[n] \overset{\text{i.i.d.}}{\sim} \text{Poisson}(A)$. A sample is affected by impulsive noise iff $m[n] > 0$, which occurs with probability

$$p_{\text{sample}} \triangleq \mathbb{P}(m[n] > 0) = 1 - e^{-A}. \tag{5}$$

Let $K \triangleq N/L$ be the number of symbols in a waveform. For each symbol index $i \in \{0, \ldots, K-1\}$, define the sample index set $\mathcal{I}_i \triangleq \{i \cdot L, i \cdot L + 1, \ldots, (i+1) \cdot L - 1\}$. A symbol is affected iff at least one of its L samples is affected. By independence across samples,

$$\begin{aligned} p_{\text{sym}} &\triangleq \mathbb{P}(\mathbb{1}\{\exists n \in \mathcal{I}_i : m[n] > 0\} = 1) = 1 - \mathbb{P}(m[n] = 0, \forall n \in \mathcal{I}_i) \\ &= 1 - (1 - p_{\text{sample}})^L = 1 - e^{-A \cdot L}. \end{aligned} \tag{6}$$

Since $\mathbb{1}\{\exists n \in \mathcal{I}_i : m[n] > 0\}$ are independent and identically distributed (i.i.d) across disjoint symbols, the number of affected symbols

$$S \triangleq \sum_{i=0}^{K-1} \mathbb{1}\{\exists n \in \mathcal{I}_i : m[n] > 0\} \sim \text{Binomial}\big(K,\, p_{\text{sym}}\big), \tag{7}$$

which yields $\mathbb{E}\big[\frac{S}{K}\big] = p_{\text{sym}}$.

Targeting a desired symbol-hit rate. To target an *average* fraction $p^\star = 0.15$ of affected symbols per waveform, set $p_{\text{sym}} = p^\star$ and solve for A:

$$A^\star = -\frac{1}{L}\, \ln(1 - p^\star) = \frac{-\ln(0.85)}{L}. \tag{8}$$

This calibration ensures $\mathbb{E}[S/K] = p^\star$. Because S is binomial, the realized fraction $\hat{p} = S/K$ fluctuates around $p^\star$ for finite K; achieving *exactly* 15% per waveform is impossible without additional (and statistically distorting) conditioning such as rejection sampling or direct symbol selection.

A.2 Concentration of the Empirical Affected-Symbol Ratio

Let $p \triangleq p_{\text{sym}}$. For any $\epsilon \in \big(0, \min\{p, 1-p\}\big)$, a standard two-sided Chernoff bound for Bernoulli means [16] gives

$$\mathbb{P}\left(|\hat{p} - p| \geq \epsilon\right) \leq 2 \exp\Big(- K D_{\text{KL}}(p + \epsilon \| p)\Big), \tag{9}$$

where, for $q, p \in (0, 1)$, the binary Kullback–Leibler divergence (KL) divergence is

$$D_{\text{KL}}(q \,\|\, p) = q \ln \frac{q}{p} + (1 - q) \ln \frac{1 - q}{1 - p}. \tag{10}$$

There is no closed-form inverse for ϵ as a function of the tail probability δ, but for small ϵ a quadratic expansion yields [17]

$$D_{\text{KL}}(p + \epsilon \,\|\, p) \approx \frac{\epsilon^2}{2\, p(1-p)} \Longrightarrow \epsilon \approx \sqrt{\frac{2\, p(1-p)}{K}\, \ln\Big(\tfrac{2}{\delta}\Big)}. \tag{11}$$

With $\delta = 0.05$, $p = 0.15$, and $K = 128$, this gives $\epsilon \approx 0.085726$, i.e.,

$$\hat{p} \in \big[p - \epsilon,\ p + \epsilon\big] \approx [0.064274, 0.235726], \tag{12}$$

with probability at least 95%.