Using LZMA Compression for Spectrum Sensing with SDR Samples

André R. Rosete
Interdisciplinary Telecom. Program
University of Colorado Boulder
Boulder, Colorado, USA
andre.rosete@colorado.edu

Kenneth R. Baker
Interdisciplinary Telecom. Program
University of Colorado Boulder
Boulder, Colorado, USA
kbaker@colorado.edu

Yao Ma
Communications Technology Laboratory
National Institute of Standards and Technology
Boulder, Colorado, USA
yao.ma@nist.gov

***Abstract*—Successful spectrum management requires reliable methods for determining whether communication signals are present in a spectrum portion of interest. The widely-used energy detection (ED), or radiometry method is only useful in determining whether a portion of radio-frequency (RF) spectrum contains energy, but not whether this energy carries structured signals such as communications. In this paper we introduce the Lempel-Ziv Markov-chain Sum Algorithm (LZMSA), a spectrum sensing algorithm (SSA) that can detect the presence of a structured signal by leveraging the Liv-Zempel-Markov chain algorithm (LZMA). LZMA is a lossless, general-purpose data compression algorithm that is widely available on many computing platforms. The new algorithm is shown to have good performance at distinguishing between samples containing communication signals, and samples of noise, collected with a software-defined radio (SDR). This algorithm does not require any information about the signal beforehand. The algorithm is tested with computer-generated as well as SDR-captured samples of LTE signals.**

***Index Terms*—spectrum sensing algorithm, compression, SDR, ROC, AUC**

I. Introduction

COMPRESSION algorithms remove redundancies in data sets to encode data using fewer bits. Communication signals, unlike noise, contain regularly reoccurring features; a reflection of the fact that they are protocols designed to convey information. While Shannon [1] has shown that the optimal channel coding theorem is a randomly-encoded signal, actual attempts at communication are pseudo-random at best because the random sequence needs to be able to be regenerated by the receiver to demodulate the signal. Since the compressibility of a sequence serves as a test of its randomness [2], we can expect that communication signal samples will be more compressible than Gaussian noise samples, as the latter are random values. The Lempel-Ziv-Markov chain Algorithm (LZMA) is a lossless, general-purpose data compression algorithm designed to achieve high compression ratios with low compute time. [3] We have found that, all else held constant, SDR-collected, LZMA-compressed files containing in-phase and quadrature (IQ) samples signals with higher signal-to-noise ratios (SNRs) show better compression ratios than files containing IQ samples of lower-SNR channels, or of channels known to contain only noise. Because a significant difference exists in the size of compressed sample sets of noise and compressed sample sets of communication signals, the compressibility of a spectrum sample set presents itself as a possible test statistic for spectrum sensing. Throughout this paper, we consider noise specifically as white Gaussian noise.

In this paper we explore the utility of this phenomenon in detecting the presence of communications signals from a sampled waveform at various SNRs. Section II sets the mathematical notation for our analysis and explains our evaluation metric. Section III summarizes LZMA's operation. Section IV introduces the new Lempel-Ziv Markov-chain Sum Algorithm (LZMSA). Section V explains how the sample sets for testing the algorithm were generated, and describes the testing procedure and results. Section VI presents discussion of the results.

II. Spectrum Sensing

In a general sense, spectrum sensing is the task of quantifying how the spectrum is occupied. This can be considered from the general occupancy, structure, or protocol point of view. The choice depends on the objectives of the spectrum sensing task. Often, spectrum sensing is used to refer to on/off detection - whether there is any kind of occupant in a channel beyond some threshold, usually set at some estimate of the noise floor. However, it may prove necessary to probe whether a present waveform has structure, which would indicate the presence of a spectrum occupier other than noise, or to take it further, identify which protocol the structured waveform follows.

This work utilized the RMACC Summit supercomputer, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University.

Certain commercial equipment, instruments, or software are identified in this paper in order to adequately specify the experimentation procedure. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the software or equipment identified are necessarily the best available for the purpose.

This work was performed under the financial assistance award 70NANB18H006 from the National Institute of Standards and Technology, U.S. Department of Commerce.

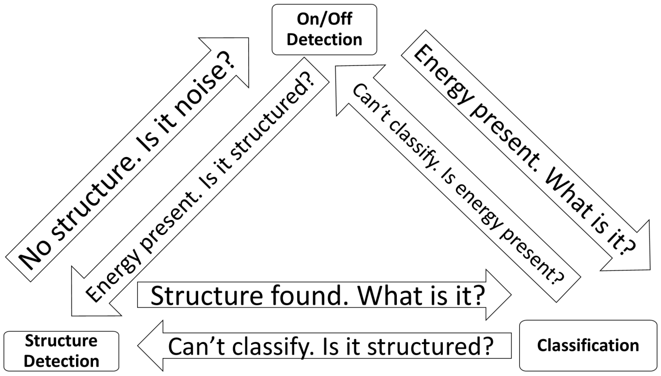
On/Off Detection
No structure. Is it noise?
Energy present. Is it structured?
Can't classify. Is energy present?
Energy present. What is it?
Structure found. What is it?
Structure Detection
Can't classify. Is it structured?
Classification

Figure 1. Spectrum Sensing Tasks.

Fig. 1 shows the flow of the spectrum sensing task, which revolves around three stages, on/off detection, structure detection, and classification. On/off detection is simply estimating whether energy is present on a threshold. Classification is identifying what kind of signal protocol is present, e.g. LTE, Wi-Fi, FM, etc. In this manuscript we focus on structure detection. Structure detection is the task of determining whether a channel $x(t)$ contains a structured signal $s(t)$ in addition to noise $n(t)$. The null hypothesis H_0 is that only noise is present in the channel, and the detection hypothesis H_1 is that a structured signal, such as a communications signal, is present in the channel in addition to noise. We use the definition:

$$\hat{H}_0, \quad \text{if } x(t) = n(t) \qquad t \in [0, T] \tag{1}$$
$$\hat{H}_1, \quad \text{if } x(t) = n(t) + s(t) \qquad t \in [0, T] \tag{2}$$

where t is a time in a continuous period of time of duration T.

A spectrum sensing algorithm (SSA) is usually implemented on digital systems which sample and quantize the signal into values that are discrete in both time and magnitude. The resulting N_s samples are denoted:

$$\boldsymbol{x}[n] = [x_0[n], x_1[n], \ldots, x_{N_s-1}[n]]^T \tag{3}$$

where each n is an integer, a discrete point in time on which each sample of the channel was taken, relative to the start of the sampling. An SSA takes $\boldsymbol{x}[n]$ as an input and returns a test statistic γ, which is a score meant to estimate the probability that $x(t)$ contains some $s(t)$. However, γ is not a measure of probability, and has different meanings depending on the SSA. For example, in ED, γ is the total energy contained in the sample set, while in covariance-based methods γ is a function of the relationship between covariance measurements across the sample set. Let us define $x[n]$ as the set of channel samples, p_{fa} the desired probability of false alarm (false positive rate), and L as the window size for covariance-based SSAs. In all cases, the decision on whether a channel contains a signal is made by comparing the test statistic with a chosen threshold γ_0. The choice of γ_0 is made to reach some desired p_{fa} based on the constant false alarm rate criterion [4].

$$\text{decision} = \begin{cases} H_0, & \text{if } \gamma \leq \gamma_0 \\ H_1, & \text{if } \gamma > \gamma_0 \end{cases} \tag{4}$$

Since we wish to compare various types of detection methods, we are interested in the performance of the SSAs independent of the choice of threshold.

One such threshold-independent evaluation method is the area under the Receiver Operating Characteristic (ROC) curve, or Area Under the Curve (AUC) [5]. The ROC curve is a plot of the true positive rate (TPR) versus the false positive rate (FPR) of an SSA.

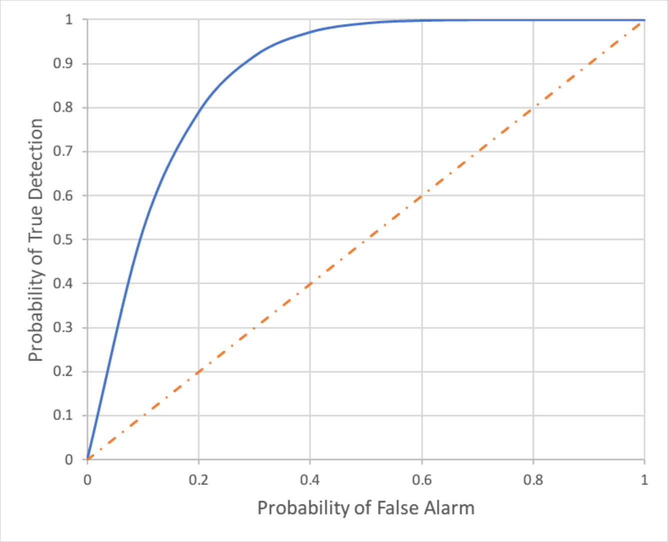

Figure 2. Receiver Operating Characteristic Example.

Fig. 2 shows a ROC curve (solid blue line) and an ignorance line (dashed orange line). Any portion of an SSA's ROC curve under the ignorance line means that the SSA is worse than random guessing at that point in the curve. AUC is the area between the ROC curve and the x-axis.

The area under the ROC curve (AUC) is a value between 0 and 1 that summarizes the statistical strength of a classifier. AUC at or below 0.5 means that the statistical strength of the classifier is no better than guessing and is therefore not useful. AUC at 1 means that the classifier always makes the correct decision and never makes an incorrect decision. AUC at 0 means that the classifier always makes an incorrect decision [4]. The benefit of AUC is that it enables a performance evaluation of spectrum-sensing algorithms by summarizing the ROC curve with a single number, which permits one to visualize performance as a function of a number of parameters, such as the number of samples or SNR, both important constraints in spectrum sensing implementations.

ED [6] is a popular SSA, with applications from the Federal Communication Commission's (FCC) spectrum regulation, where interference thresholds are measured in microvolts per meter [7], to Wi-Fi coexistence [8]. It works by calculating how much energy is contained in the received samples $x[n]$, working on the assumption that any electromagnetic emission contains energy. This can be a disadvantage when specifically looking for communications signals, since ED will detect non-communications emissions as well. However, ED may be particularly useful in detecting natural interferers or to get a quick overview of which channels may warrant further scrutiny with more structure-aware SSAs. ED is not a communications detector, as it will detect natural noise, based on its energy content, just as much as a communications signal at the same power.

One example of an SSA meant to look for structure in a signal, included here for comparison, is the Covariance Absolute Value (CAV) algorithm described in [9] in Eqs. 26-30. This SSA seeks to measure the self-covariance of a set of samples within a window of l samples, with the assumption that samples of a structured signal tend to be more co-variant than samples of noise. Like LZMSA, this algorithm does not require any information about the signal beforehand.

III. Lempel-Ziv-Markov Chain Algorithm

In order to understand how the LZMA algorithm might be useful detecting signals within noise, it is instructive to understand the basics of how this compression algorithm operates. LZMA is a chain of three compression algorithms: 1) a delta encoder, 2) the compression algorithm known as "LZ77", and 3) a range encoder [10].

Initially, the input data is processed by the delta encoder as follows [11]:

$$\delta(v_1, v_2) = (v_1 \setminus v_2) \cup (v_2 \setminus v_1) \tag{5}$$

where v_1 represents the first sequence in the sliding window, and v_2 represents the second sequence in the sliding window, $\setminus$ is the set minus operator, and $\cup$ is the set union operator. For example, the data set [3, 4, 6, 9, 3] would be stored as [3, 1, 2, 3, -6]; each data point in the set is stored as the difference from the previous data point in the set, with the exception of the first value, which is kept as-is; it is the difference from zero.

The resulting sequence is input to the LZ77 compression algorithm [12]. This resulting sequence is then input to a range encoder [13].

The purpose of this chain operation is that the delta encoder prepares the data in such a way that it may be more compressible by the sliding window algorithm in LZ77. The output is then passed on to a range encoder, which is able to further remove some redundancies not caught by the LZ77 algorithm.

The goal of LZMA is to compress data as much as possible with a reasonably low processing time [3]. Furthermore, a fast field-programmable gate array (FPGA) implementation of LZMA has been demonstrated [10].

IV. LZMSA Detection

We propose a new signal detection algorithm based on LZMA. The algorithm begins by compressing the set of samples $\boldsymbol{x}[n]$ with the LZMA compression algorithm.

$$\boldsymbol{x}[n] \xrightarrow{\text{LZMA}} \boldsymbol{y}[m] \tag{6}$$

Let $\boldsymbol{y}[m]$ be the LZMA output for $\boldsymbol{x}[n]$. $\boldsymbol{y}[m]$ is a data object containing M bytes.

$$\boldsymbol{y}[m] = [y_0[m], y_1[m], \ldots, y_{M-1}[m]]^T \tag{7}$$

The LZMSA decision statistic is produced as follows:

$$\gamma_{LZMA} = \underline{1}^T \boldsymbol{y}[m] = \sum_{k=0}^{M-1} y_k[m] \tag{8}$$

where $\underline{1} = [1, \ldots, 1]^T$.

One possible explanation for why this method may work is that information-carrying signals tend to be more compressible by LZMA than samples representing only random noise. By the nature of random numbers, it is a reasonable expectation that a sequence of random numbers cannot be significantly compressed [2]. This is due to the fact that truly random numbers, such as those in Gaussian noise, already represent a sequence composed almost entirely of Shannon information, which would represent maximum entropy. The Kolmogorov complexity [14] of such a random sequence, that is, the shortest possible descriptor that could fully generate such a sequence, cannot be shorter than the sequence itself.

V. Test Sample Sets Generation and Testing

Two data sets were generated: A) A digital chain set generated with MATLAB™ R2017b running on the Rocky Mountain Advanced Computer Consortium (RMACC) Summit supercomputer. B) A set generated with a vector signal transceiver (VST), transmitted over a coaxial copper cable, and captured with a SDR. In both cases, the target SNRs for the sample sets were -12 dB through 12 dB in increments of 3 dB.

A. Computer-Generated Signal Set

Digital chain samples were generated in MATLAB™ , forming a fully computer-generated (CG) set, free of potential hardware artifacts. These samples were of Long-Term Evolution (LTE) Test Model 3.3, with a bandwidth of 20 MHz. The samples were summed with generated Gaussian noise to reach the target SNRs.

B. SDR Set

LTE Test Model 3.3 signals were generated with a National Instruments™ PXIe-5645R Vector Signal Transceiver. The signals were captured using an Ettus Research™ B200mini Software-Defined Radio (SDR) controlled by a PC running GNU Radio on Ubuntu Linux. All signals were generated and captured at a center frequency of 5.8 GHz, which is a relevant spectrum portion due to the fact it is shared between Band 255 LTE and U-NII-3-band Wi-Fi. The SDR captured 20 MHz of bandwidth at 20 million samples per second, which is the Shannon rate considering that each IQ sample consists of both an in-phase and a quadrature component. Signals were generated at different power levels to reach the target SNRs with respect to the hardware noise. Both the LTE and noise samples collected in this manner were found to be non-zero mean. Furthermore, the noise samples were found to show IQ correlation even though noise samples should be expected to be independent.

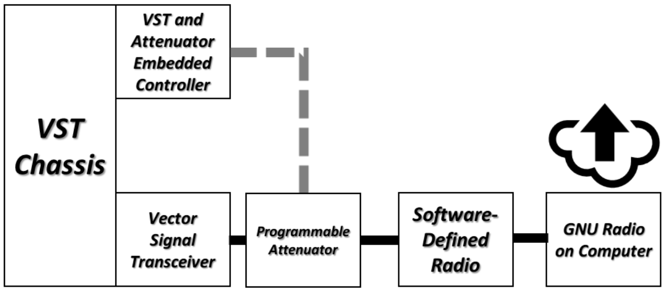

Figure 3. VST-SDR Generation and Capture.

Computations were carried out using a Python™ program to compare the detection performance of LZMSA versus the CAV algorithm on the Gaussian and LTE sample sets. Simulations were run for each SNR set, with N_s from 10 to 5000 samples per decision in steps of 100 samples. 1000 Monte Carlo trials per SNR and number of samples were performed to calculate the AUC. The covariance window was set at 10 samples for the CAV algorithm. All tests were run on the RMACC Summit supercomputer. The simulations consist of Monte Carlo trials where, for each number of samples N_s and and each SNR, the algorithm under test is fed a randomly selected consecutive set of N_s samples from the available signal samples file, repeating the process for the noise-only samples file. Test statistics are produced for both detection hypotheses, H_1 and H_0. These test statistics are processed through an algorithm [15] which produces the AUC measure for the particular scenario, indicating how much the test statistics tend to be different between H_1 and H_0 cases.

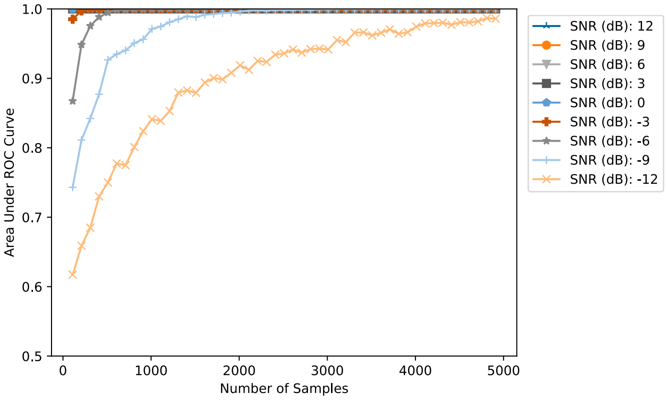

Figure 4. Energy Detection on a Gaussian Emitter.

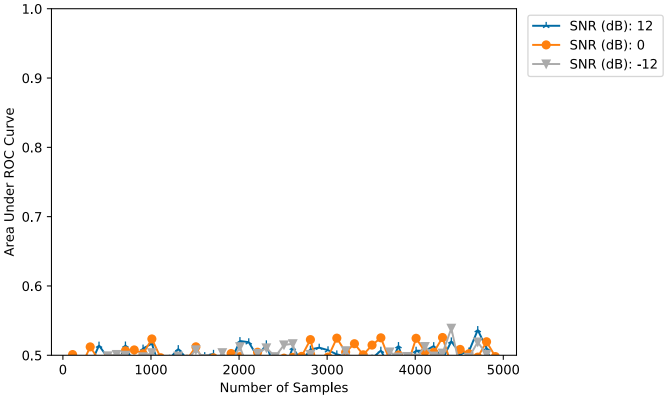

Figure 5. Covariance Absolute Value Detection on a Gaussian Emitter.

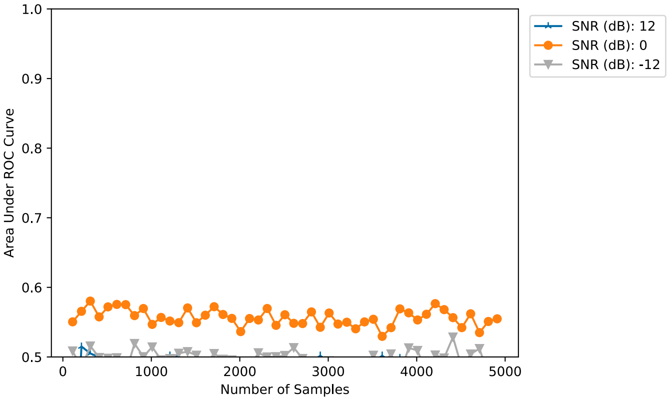

Figure 6. LZMSA Detection on a Gaussian Emitter.

In Figs. 4, 5, and 6, a Gaussian emitter was added to the background noise and compared to when only background noise was present. Fig. 4 shows that ED can yield high AUCs, even when SNR is low or when the number of samples is under 1000. AUC is strongly dependent on SNR and number of samples. The generally high AUC means that ED differentiates well between a Gaussian emitter, known to lack communications, and the background Gaussian noise. This shows that ED cannot differentiate communications from a Gaussian emitter. A communications detector should have AUC close to 0.5 in this case, as both the emitter and the background channel are lacking in communications content, and are thus indistinguishable to such a detector. Figs. 5 and 6 show that CAV and LZMSA both are incapable of differentiating between a Gaussian emitter and background Gaussian noise, regardless of the SNR or number of samples. This is due to the fact that both the emitter and the background noise lack features that would be found in communications signals, such as sample covariance or redundancy.

Fig. 7 shows CAV consistently outperforms LZMSA when CG samples are used, with CAV's AUC increasing with number of samples, when SNR is 12 dB or 0 dB. However, when SNR is reduced to -12 dB, both algorithms have an AUC around 0.5 regardless of the number of samples used.

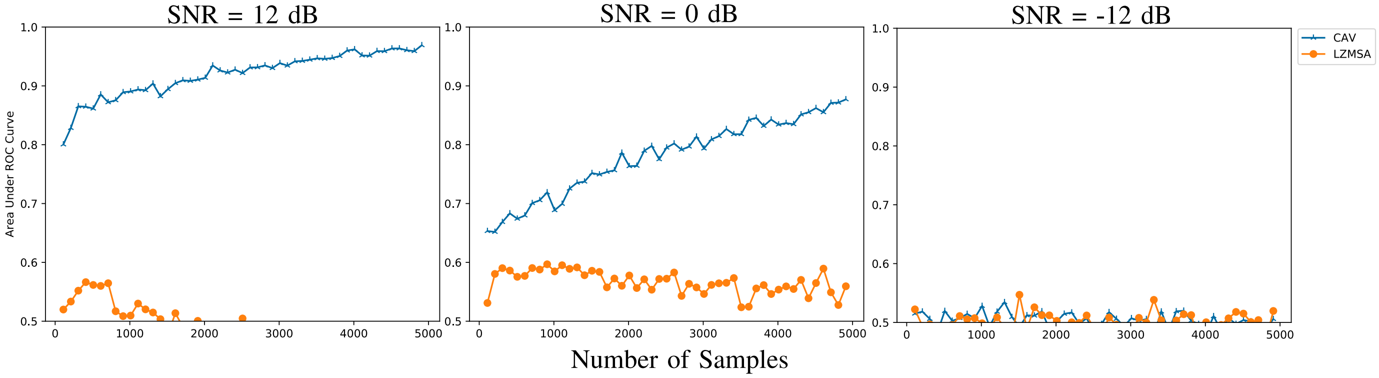

Figure 7. CAV vs. LZMSA on CG LTE Signal.

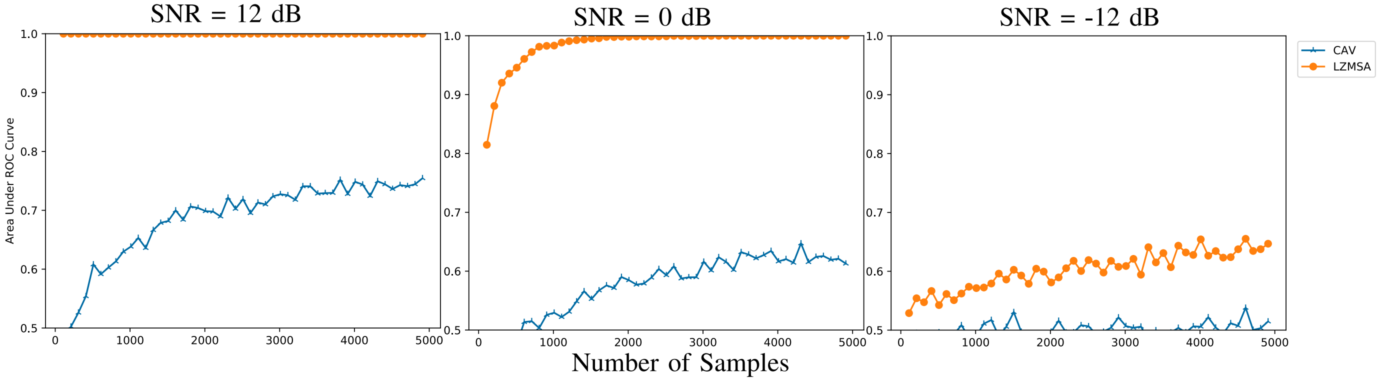

Figure 8. CAV vs. LZMSA on VST-SDR LTE Signal.

Fig. 8 shows LZMSA consistently outperforming CAV on all three SNRs when SDR-collected samples are used, with AUC increasing as more samples are used. Even with an SNR of 12 dB, CAV shows consistently low performance; at 5000 samples, its AUC is around 0.75, while LZMSA's AUC is 1 even at 10 samples. At 0 dB and 2000 samples, CAV's AUC is under 0.6, while LZMSA's is 1. Not only is the situation reversed from where the CG LTE samples were used, but LZMSA actually shows higher AUC across the board for VST-SDR samples than CAV does for the CG LTE samples.

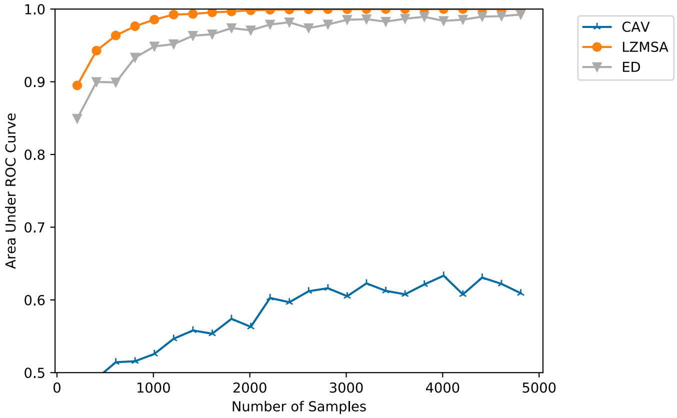

Figure 9. ED vs. LZMSA on VST-SDR LTE Signal. SNR = 0 dB.

Fig. 9 shows that there are some cases where LZMSA even performs better than ED for classical on/off detection. LZMSA shows a higher AUC than ED, with a greater gap in AUC evident when using fewer samples. Both outperform CAV significantly in this case, with CAV's AUC barely rising above 0.6 when N_s is above 3000. AUC shows little positive relation to the number of samples used.

The SDR-collected samples of both LTE and noise were found to be non-zero mean and IQ-correlated, which may offer insight into the wide performance discrepancies found between the CG and VST-SDR scenarios. It could be that the IQ correlation present in the VST-SDR noise samples made it difficult for the CAV algorithm to differentiate between LTE and noise, since the mechanism through which this algorithm differentiates a structured signal from noise is the difference in sample correlation, with the assumption that noise samples are independent and uncorrelated. If correlation is introduced to the noise samples, they cannot be well differentiated from a correlated waveform like LTE on this basis. However, it is unclear why these effects enable LZMSA to work so well with the VST-SDR samples as opposed to the CG samples.

VI. Conclusion

In this manuscript we tested three spectrum sensing algorithms, ED, CAV, and LZMSA, in a scenario where both the emitted signal and the background noise are Gaussian. We showed ED to not be a communications detector because its high AUC indicated differentiation between a Gaussian emitter and the background Gaussian noise. We then showed CAV and LZMSA to not be sensitive to this Gaussian emitter. The next scenarios were comprised of CG and SDR-implemented LTE Test Model 3.3 signals where we compared the performance of CAV and LZMSA. While CAV is shown to differentiate CG LTE samples from the background noise fairly well, this result doesn't apply to the VST-SDR LTE samples, where performance is poor. Conversely, while LZMSA shows poor performance on CG LTE samples, we show that it can perform well on SDR-collected LTE samples. The causes for the discrepancy in the performance of these two algorithms between the CG and SDR-collected scenarios warrant further research into the interactions between spectrum sensing algorithms and effects introduced by SDRs on waveforms during their digitization process. Further research must also be carried out to find out how the SDR's hardware effects enable LZMSA to differentiate an LTE input from a noise input. It is of note, that LZMSA shows promise as a viable SSA that could perform well in some scenarios. Notably, it works well in conjunction with SDR collected-samples, even outperforming ED at on/off detection of an LTE signal. In our future work we will address the differentiation in performance of the algorithms between CG and SDR-collected samples. Furthermore, we will be studying the efficiency of these algorithms in the context of a spectrum monitoring application utilizing SDRs, with additional scenarios such as detection of a variety of signals in noisy conditions.

Acknowledgments

We would like to thank Dr. Adam Wunderlich for his advice on ROC and AUC theory, and to W. Max Lees for providing the fast ROC and AUC implementations in Python™.

References

[1] C. E. Shannon, "Communication in the presence of noise", *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[2] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, and S. Vo, "A statistical test suite for random and pseudorandom number generators for cryptographic applications", Booz Allen & Hamilton, NIST Special Publication 800-22, May 15, 2001.

[3] I. Pavlov, *7z Format*, 2018. [Online]. Available: https://www.7-zip.org/7z.html.

[4] H. L. Van Trees, *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons, 2004.

[5] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve.", *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1, 1982.

[6] H. Urkowitz, "Energy detection of unknown deterministic signals", *Proceedings of the IEEE*, vol. 55, no. 4, pp. 523–531, Apr. 1967.

[7] "47 CFR § 15.209 Radiated emission limits; general requirements.", in *United States Code of Federal Regulations*.

[8] IEEE, *802.11-1997: IEEE standard for wireless LAN medium access control (MAC) and physical layer (PHY) specifications.* 1997.

[9] Y. Zeng and Y. C. Liang, "Covariance Based Signal Detections for Cognitive Radio", in *2007 2nd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, Apr. 2007, pp. 202–207.

[10] E. J. Leavline and D. A. A. Gnana Singh, *Hardware Implementation of LZMA Data Compression Algorithm*, Mar. 2013.

[11] R. Conradi and B. Westfechtel, "Version models for software configuration management", *ACM Computing Surveys (CSUR)*, vol. 30, no. 2, pp. 232–282, 1998.

[12] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression", *IEEE Transactions on Information Theory*, vol. 23, no. 3, pp. 337–343, May 1977.

[13] G. Martin, "Range encoding : An algorithm for removing redundancy from a digitised message", *Video and Data Recording Conference, Southampton, 1979*, pp. 24–27, 1979.

[14] A. N. Kolmogorov, "On Tables of Random Numbers", *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, vol. 25, no. 4, pp. 369–376, 1963.

[15] X. Sun and W. Xu, "Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves", *IEEE Signal Processing Letters*, vol. 21, no. 11, pp. 1389–1393, Nov. 2014.