Identification of Frequency Modulated Signals in the Allen Telescope Array Data

Austin Hou [1], Rafael Setra [1], Qi Ian Yang [2]
mentored by Andreas Paepcke [3]
(8 June 2016)
[1] *Department of Electrical Engineering*
[2] *Department of Physics*
[3] *Department of Computer Science*

Abstract

Archival data from the Allen Telescope Array were analyzed with a variety of techniques, including Fourier transformation, fluctuation analysis, and Fisher vector calculations. 1.6±0.4% of the data records were found to contain frequency modulated signals (“squiggles”). We report a combined best-case misclassification rate of 1.5% false negative and 0.2% false positive.

1. Introduction

As a part of its quest to search for extraterrestrial intelligence, SETI Institute employs the Allen Telescope Array (ATA) to survey radio transmissions from outer space aiming to discover intelligent communication of an extraterrestrial origin. The ATA has been recording data for the past few decades, however, the collected data suffer from low signal-to-noise ratio (SNR), and discernible signals are often resulted from human-originated radio frequency interferences (RFIs), such as electromagnetic waves generated by radars and aircrafts.[1] Classification and filtration are therefore crucial initial steps to distinguish signals from unknown sources from known RFIs, in order to gain further insights into the ATA data.

The ATA consists of 42 antenna dishes that produce around 60 gigabits of data per second, in the form of complex voltage readings as functions of time (CompAmps). To visualize, a CompAmp file is often rendered as a spectrogram (energy density as a function of time and frequency, also known as “waterfall plots”; fig. 1.1) by slicing it into one-second fragments and Fourier-transform each fragment. Due to limitations in computing power in the past, the majority of the ATA data from the last ten years have been archived but largely unanalyzed. Among the features of interest found by human examination of the spectrograms, a particular loosely-defined category is referred to as “squiggles” (fig. 1.2), which resemble frequency modulated signals with the central frequencies evolve back and forth through time.

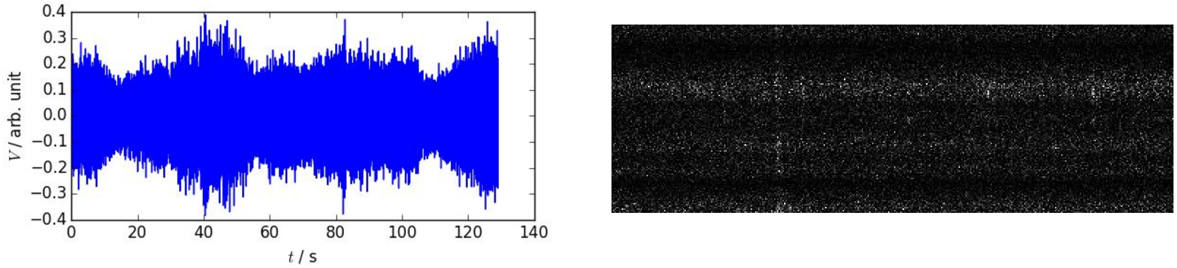

Figure 1.1. Segments of a CompAmp (left, real part only) and a spectrogram (right).

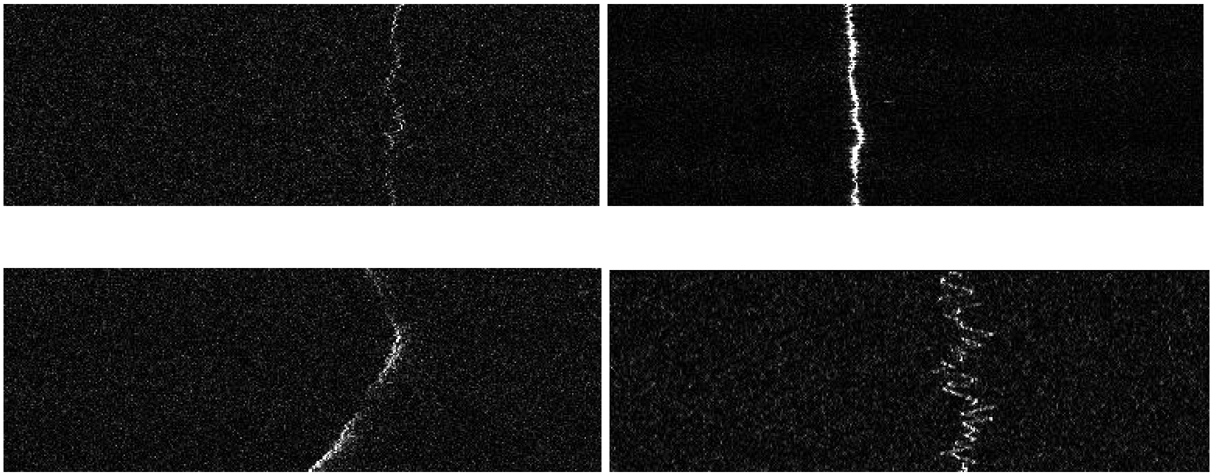

Figure 1.2. Examples of frequency modulated signals (“squiggles”).

The ATA data available to us include roughly 400,000 CompAmp records on the IBM Bluemix server, each one megabyte in size; and 10,000 pre-rendered spectrogram files, each with 1.6 million pixels, available both locally and on the IBM Bluemix server. Additionally, a collection of 800 manually picked squiggles are available as reference. Current methods of identifying squiggles by human screening is insufficient given the large size of the archival database. In this report, we summarize a project aiming to expedite the detection of such frequency modulated signals (squiggles) by applying distributed-style data-mining techniques combined with mathematical analysis and signal / image processing. The methods used to analyze the data and their respective results are described in Section 2. The overall performance of these methods are summarized and discussed in Section 3.

2. Methods and Results

In the following subsections, we present three distinct methods in identifying squiggles. Since the overall purpose is to identify as many as possible interesting features as candidate for further studies, these methods are designed to produce as low as possible false negative rates. While low false negative rates are crucial in order not to miss important discoveries, a reasonable false positive rates is a secondary objective to make further studies humanly possible. Firstly, a quick and approximate approach based on Fourier analysis and autocorrelation is outlined in Section 2.i. The main purpose of such method is to estimate the overall proportion of the data records that contain squiggles. Two additional methods were developed to further refine the misclassification rates. A second approach based on fluctuation analysis of time series is discussed in Section 2.ii. Finally in Section 2.iii, we present a clustering scheme combining Fisher vector calculation with existing SETI clustering proposals.

To train and evaluate the methods in Sections 2.ii–iii, the 10,000 locally available spectrograms were manually sampled and filtered. Roughly 300 squiggles and 7000 non-squiggle features were found in addition to the 800 hand-picked squiggles. Together these form a local data set.

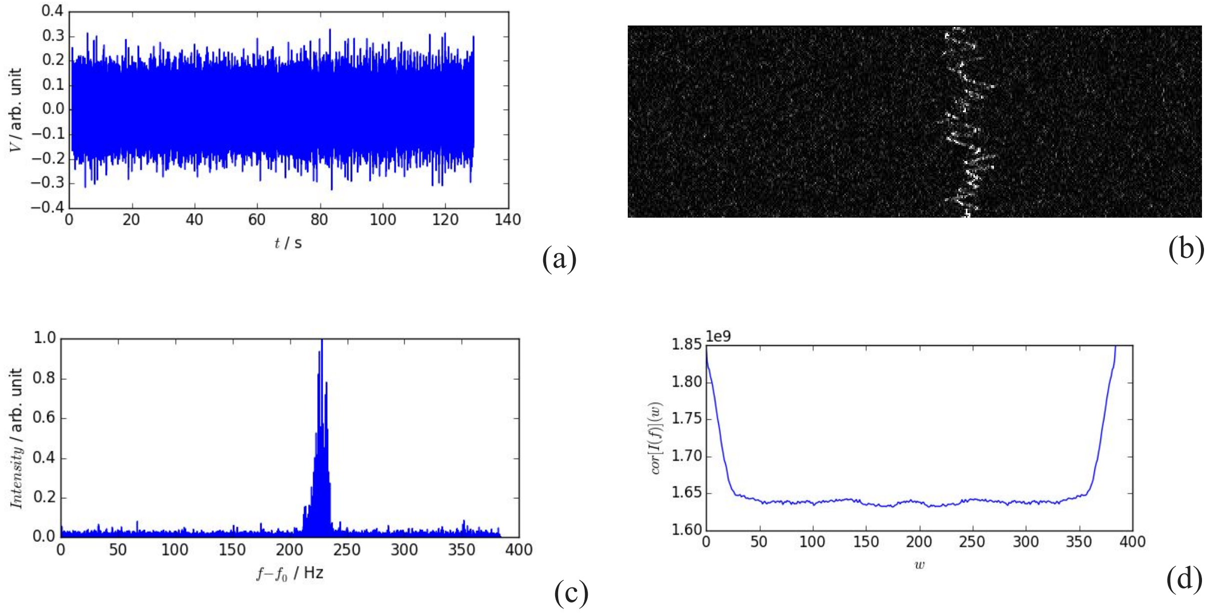

Figure 2.1.1. A squiggle represented by (a) a CompAmp, (b) a spectrogram, (c) the Fourier spectrum of its CompAmp. The autocorrelation (d) of the Fourier spectrum shows a slow decay near zero.

2.i. Fourier Analysis and Autocorrelation

From an *ab initio* perspective, a squiggle is a narrow band in the frequency domain whose central frequency evolves back and forth with time in a range significantly greater than its bandwidth. Therefore a broad peak in the frequency spectrum is expected when a time series containing a squiggle is Fourier transformed as a whole and plotted as frequency domain intensities (fig. 2.1.1). In practice, since features smaller than the narrow bandwidth is not required, the Fourier spectrum may be obtained by utilizing the existing routine of generating the spectrogram and summing over the time axis. (While the mathematical proof is outside the scope of this project, an argument can be made by considering a thought experiment with electromagnetic waves, band-pass filters and invoking the conservation of energy.)

The existence of one or multiple such broad peaks in a Fourier spectrum leads to a slow decay near the zero of the autocorrelation (with periodic boundary condition) of the Fourier spectrum (fig. 2.1.1d). Since the intensity is always real and positive, the autocorrelation is simply:

$$cor[I(f)](w) = \oint I(f)I(f-w)df$$

To quantify such slow decay, we used a simple criterion:

$$cor[I(f)](width) > threshold \cdot min(cor[I(f)])$$

where the minimum value of the autocorrelation corresponds to the overall background level in the spectrum.

The values of the parameters *width* and *threshold* in the criterion were determined by performing an exhaustive search on hand-picked squiggle samples (fig. 2.1.2). Half of the samples were used in the optimization whereas the remaining half were reserved for validation. Smaller values in either of the parameters in general result in more tolerant criteria. Consequently the false negative rates is reduced quasi-monotonously by decreasing either of the parameters. However, more tolerant criteria likely yield higher false positive rates. By selecting a trade-off point where further reducing either of the parameters no longer significantly reduce the false negative rate, the optimal parameters are chosen to be *width* = 15 and *threshold* = 1.0027, at which a false negative rate of 2% were obtained on both the training set and the validation set.

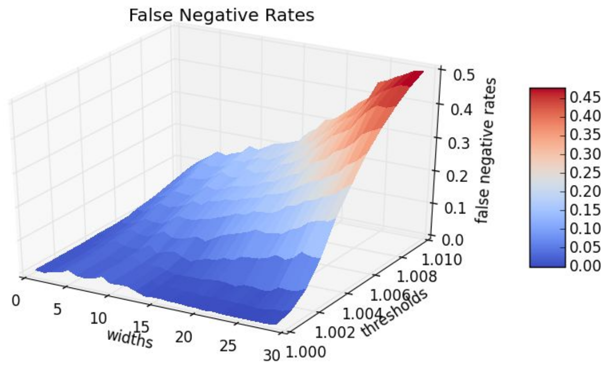

Figure 2.1.2. False negative rates as a function of the parameters *width* and *threshold*.

The values of the optimal parameters seem to indicate that the majority of the squiggle samples have their central frequency evolving within ±15 pixels, and that the faintest squiggle might be barely recognizable from the background. By deploying the routine described above on the IBM Spark system and analyzing all 400,000 archival CompAmp records, roughly 20,000 were found to exhibit broad peaks in their Fourier spectra. By visually examining 40 random samples of the resultant spectrograms, 68% of the records that satisfy the criterion were found to be false positive produced by broad vertical features other than squiggles. Assuming applicability of the central limit theorem, we estimate that overall 1.6±0.4% of the CompAmp records contain squiggle signals.

2.ii. Fluctuation Analysis

It was noted that the vast majority of squiggle examples previously found had similar qualitative features to a traditional random walk. These signals are continuous, often centered around a changing carrier, and fluctuated with some regularity. Hurst analysis is a method often used to characterize random walks (e.g. [4]):

$$E[x(t+d)-x(t)] \sim d^{H}$$

Due to the low resolution in time of the data, instead of building a full model, values of *d* ranging between 1 and 6 were used to characterize the signal and variance was used instead of expectation:

$$var[x(t+d)-x(t)] \text{ for } d = 1,2,3,4,5,6$$

For linear signals, this measure is expected to remain approximately constant as *d* changes. The same is true for signals with high noise. For all other signals, such as squiggles, the change in variance as *d* increases will characterize the magnitude of fluctuation. For most of the signals analyzed in this project, the variance is expected to “flatten” as *d* increases - as such, a maximum *d* of 6 was deemed sufficient to characterize the quality of fluctuation. The six variances were sent through a linear fit and the slope extracted.

For narrowband interference, the slope was found to be on the order of 0.01. For squiggles, values were approximately on the order of 0.1. For wideband interference, values were approximately on the order of 1 or greater. Values for various types of signals can be seen in Figure 2.2.1.

This routine based on Hurst exponents was tested on both the 800 hand-picked squiggles as well as subsets of the larger local training set described earlier in Section 2. A major problem encountered was the limitations of the semi-naive algorithm to convert a spectrogram to a time series, which was essentially a seeded local maximum tracing. The column that has the greatest average intensity was set as the starting point. From the starting point, the rows were iterated through and the local maximum will be selected within a neighborhood of the last selected maximum. When encountering noisy signals, this algorithm can potentially be thrown off the track.

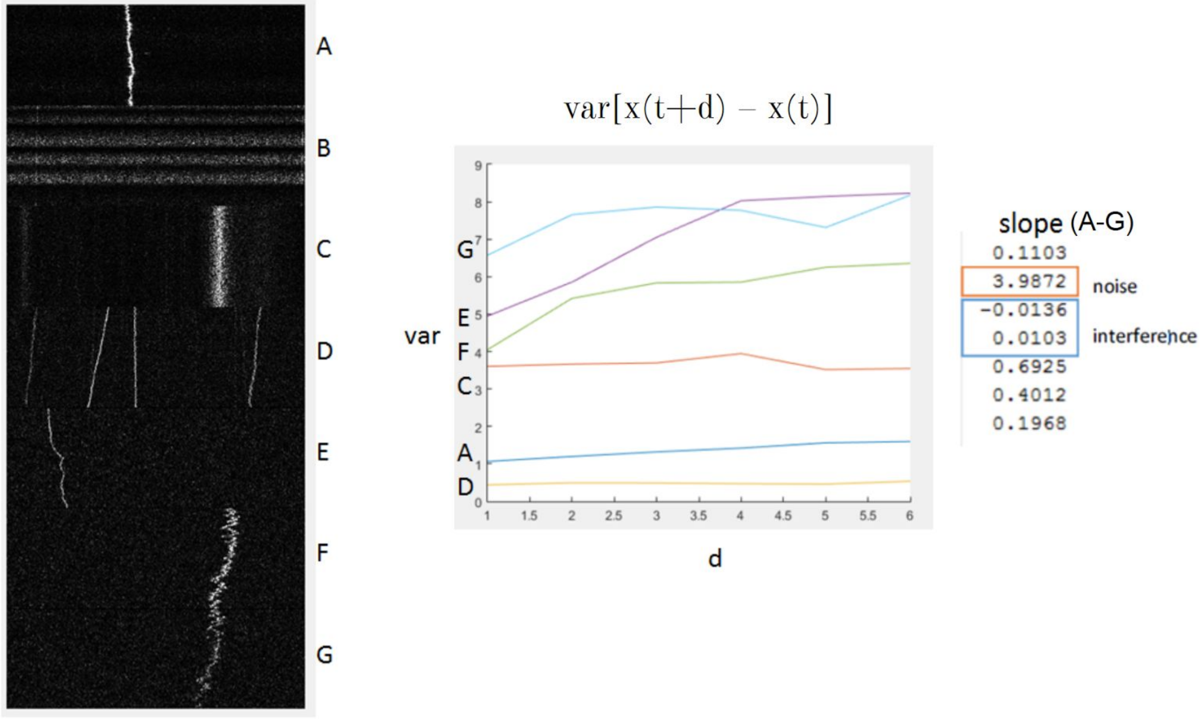

Figure 2.2.1. Signals (A)-(F) on the left side. (A),(E),(F),(G) match the characteristics of traditional squiggles identified previously. Variances for *d* = 1–6 plotted in the middle. Signal (B) omitted from the plot due to magnitude. Slope of the linear fit for each signal, respectively, on the right.

When testing, the slope for each candidate is calculated and compared against a bound. When testing for false positives, the lower bound was retained at 0.1 against the 833 hand-picked squiggles. 12 of the 833 were misclassified, resulting in a false positive rate of 1.5%. To test for false negatives, subsets of 500 out of the larger local data set were taken. A twofold approach was used, in which preprocessing was first used to eliminate all signals with a total (2D sum) of less than a threshold rate to eliminate noise. In the second phase, the bounds were applied against the signals. It was found that this false positive rate was highly dependent on the signal tracing algorithm used, and tweaking the parameters of the tracing algorithm had significant effects on the output. False positive rates of between 10-60% were found for the various subsets, with an average of 25%. We hypothesize that by using a more accurate tracing algorithm, these numbers could be improved significantly.

2.iii. Fisher Vectors and Clustering

While the two methods above are traditional methods for processing generic signals, the approach outlined in this subsection is a combination of machine learning and image processing techniques. Firstly, partly based on previous work done in NASA and SETI [5, 6], scalar features were extracted from the spectrogram images. The features used were: the standard deviations of the row and column sums, the autocorrelation width, the total energy (total sum), the Shannon entropy, and the total variation. In addition to these scalar features, Fisher vectors were used to represent key global image features. Fisher vectors are calculated in three steps. The first step is to find the scalar-invariant feature transform (SIFT) features of the image [2]. Approximately speaking, these features are representations of corners, edges, and curves in an image (fig. 2.3.1). Then the SIFT features of all of the images are collected together to form a gaussian mixture model by least likelihood. The final step is to then compare the SIFT images of one particular image to the model, and the residuals are referred to as Fisher vectors [3].

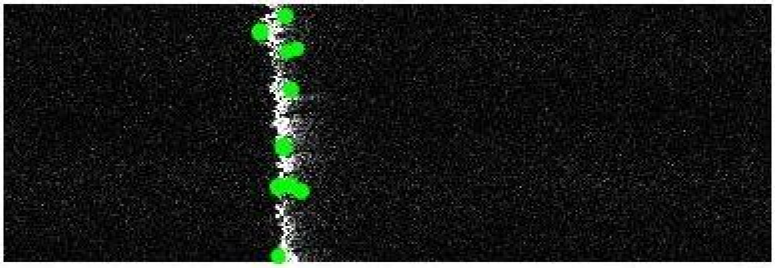

Figure 2.3.1. Example of SIFT features (green circles) in a spectrogram.

The actual classification method itself aggregates the scalar features and the Fisher vector together into one larger feature vector. The 10,000 locally available spectrogram samples are sliced into segments of the same size as the hand-picked squiggle samples. By using k-means++, the squiggles are clustered into one cluster using these vectors, whereas the non-squiggle spectrogram segments are clustered into 10 additional clusters. This number of clusters was chosen based on empirical results; it allowed for the lowest false negative rate. For a new input spectrogram, we calculate its feature vector, and choose the cluster by closest Cartesian distance. The classification scheme was trained a random half of the training data, and then validated on the remaining half. In the spirit of bagging, 10 such random trials were carried out, and the average misclassification rates were found to be 2.0% false negative and 0.2% false positive.

3. Conclusion and Discussions

To summarize, we report that 1.6±0.4% of the 400,000 archival CompAmp records contain squiggle-type signals. The misclassification rates of the three methods discussed above are displayed in Table 1. All three methods have achieved the primary objective of low false negative rates. Among these methods, the clustering scheme with Fisher vectors offers the best false positive rates. In comparison, the first two methods yielded acceptable but considerably higher false positive rates. On the other hand, the two mathematics-based methods produced classification criteria which can be easily interpreted by human, whereas the significance of the clustering scheme remains somewhat opaque.

	Fourier Analysis & Autocorrelation	Fluctuation Analysis	Fisher Vectors & Clustering
False Negative Rate (%)	2.0	1.5	2.0
False Positive Rate (%)	68	~25	0.2

Table 1. Misclassification rates of the three methods.

The methods we presented in this report are undoubtedly significant improvements over human screening. Potential future work might involve adaptation to a real-time algorithm that can be deployed in the day-to-day operations of the ATA, or to combine the three methods with an ensemble scheme such as a neuron network. Furthermore, the dynamic programming algorithm to find continuous paths of local maxima that the other SETI team in this class developed [7] might significantly reduce the false positive rates in the first two methods discussed here, without sacrificing their mathematical interpretability.

Acknowledgements

We would like to thank Dr. Jeffrey Scargle and Dr. Chris Henze from NASA, Dr. Jill Tarter and Dr. Gerry Harp from the SETI institute for insightful discussions. We are especially grateful to Graham Mackintosh from IBM, Prof. Jeffrey Ullman and all members of the teaching staff for making this project possible.

References

[1] A. Krizhevsky, I. Sutskever, & G. E Hinton, *NIPS* (2012)

[2] D. G. Lowe, *IJCV* (2004)

[3] J. Sanchez, F. Perronnin, T. Mensink, & J. Verbeek, *IJCV* (2013)

[4] A. Carbone, G. Castelli, H. E. Stanley, *Physica A*: **344**, 1–2, (2014), p. 267-271

[5] J. Scargle, *Notes on Analysis of Allen Telescope Array Data*, unpublished

[6] G. Mackintosh, *IBM jStart Team Meeting*, unpublished
[7] F. Fan, K. Smith, J. Wang, unpublished