Environmental Noise Contamination Detector – Data Pipeline

Todd Schultz, Sean Miller, Rahul Birmiwal

Due: 28 Nov 2018

Problem Statement

Aircraft certification requires extensive testing including flyover noise measurements. The measurements are acquired in remote locations to minimize contamination from environmental noise that may alter the recorded noise levels that are then submitted to regulatory agencies such as the FAA. Contamination may also cause costly repeat flyovers to acquire clean recordings for the certification process. Typical sources of contamination include bird chirps, other wildlife or livestock vocalizations, insect noise, and traffic noise. The current process involves using three engineers to listen to the live microphone feeds and alert the test manager of any noise contamination. The goal of this project is to study the feasibility of machine learning algorithms performing the task of identifying noise contamination. To understand this, our project will include a wide survey of feature sets and classification algorithms to rank the combinations and effectiveness of each.

Figure 1: A common bird found in Montana, the western meadowlark. (By Kevin Cole CC-BY 2.0)

Data source

The data used for this project originates from files. Typically, wave binary files, but the input data sources are not restricted to only this type of file. Since this is a research and development project,

all data is expected to be in file format instead of streams or API calls. Additional sample data may be required to expand the examples of wildlife noises.

The data set consists of 67 professionally recorded and labeled audio files near a major international airport. The data were recorded on three separate days in the months of November 2017, December 2017, and January 2018. The data consist of 49 aircraft recordings, 7 wildlife noise recordings, and 10 ambient recordings and vary in length from 10 seconds to a couple of minutes. Each recording was captured with scientific instrumentation grade equipment at a sampling frequency of 51.2 kHz and with a usable bandwidth from 4.6 Hz to 20 kHz. No missing samples or other anomalies are present in the data; however, the recorded levels have been normalized to allow for public release. This should not impact the project as the levels are to be normalized to remove the effects of differences in propagation distances from sound sources to the microphone sensors. Examples for the time series recordings for a bird chirp and a jet aircraft are shown in Figure 2 and Figure 3 respectively. Figure 4 shows the installation of the microphone in the field with a hemispherical wind screen to mitigate wind noise.

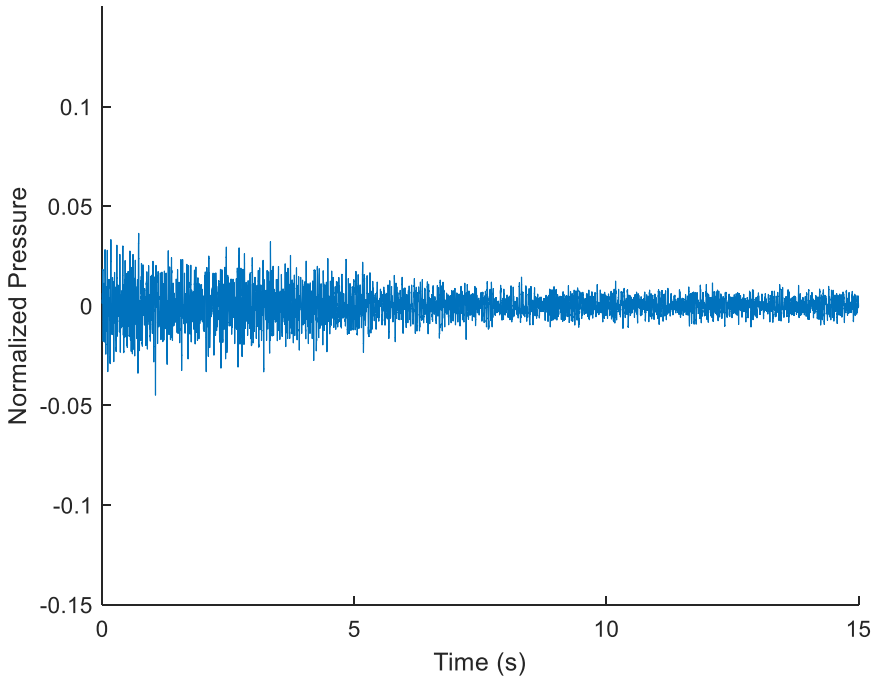

Figure 2: Example of the time series data for a bird chirp. (From file bird01.wav.)

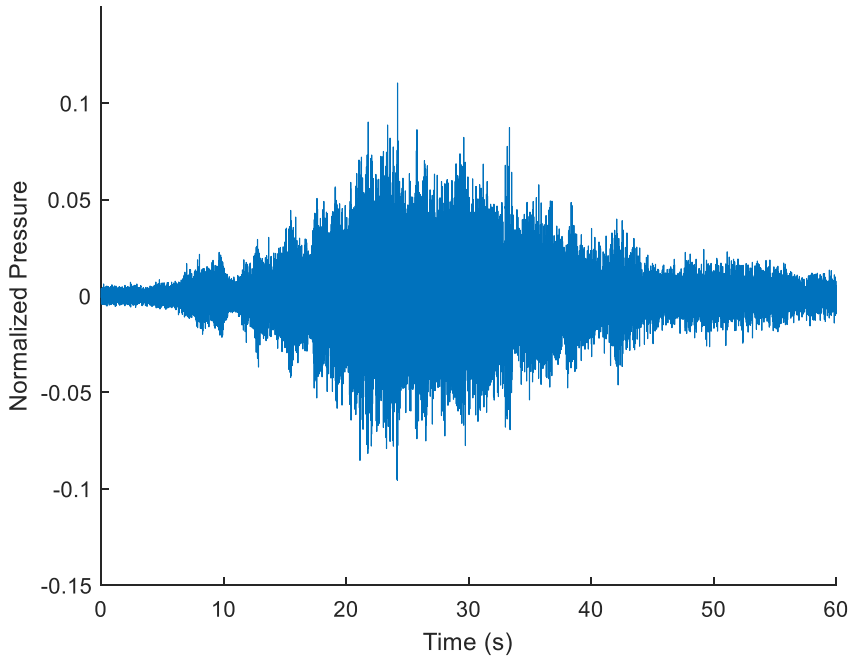

Figure 3: Example of the time series data for a jet aircraft. (From file plane08.wav.)

The signals will be broken down into smaller increments typically on the order of one to two seconds, thus each file will generate numerous feature sets. The files have been provided by the project contact to each team member to store locally as total size is over 500 MB and not suitable for storage in a GitHub repository.

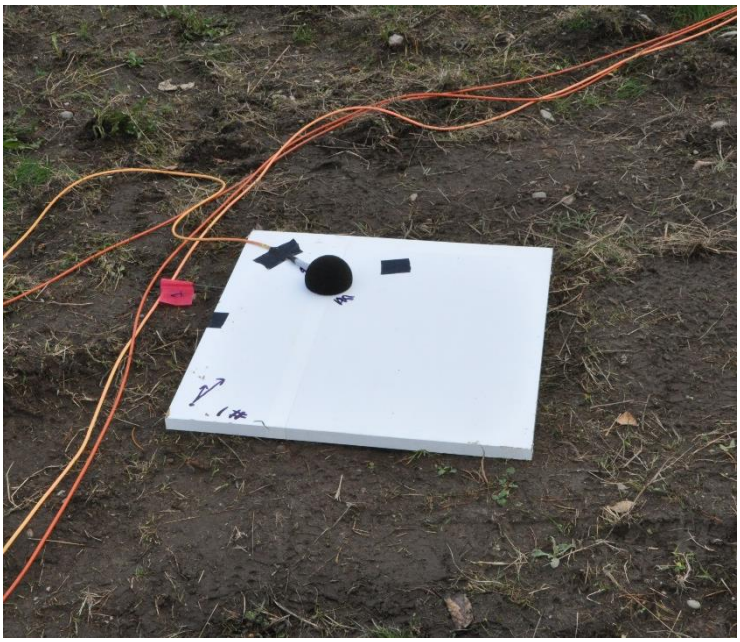

Figure 4: Microphone installation.

Data issues

Despite the data being of high quality, there are concerns with its use in this project. The recorded data is heavily skewed towards aircraft signatures thus creating an imbalance in the number of samples for each class. Additionally, each signal can produce 50-100 blocks or samples of features for classification and each block may contain 10's to 1,000's of features depending on the how features are generated. This will present a data management concern as all the data, features, and classification labels must be accurately tracked through the data processing. The class imbalance issue can be addressed by creating additional data by combining a clean aircraft signature with a contamination signal. This has the added benefit of being able to control and study the impact that the signal-to-noise ratio has on the system performance. If these combinations aren't sufficient, public domain recordings are available for use to extend the set of wildlife/livestock vocalizations included in this study.

Data example

Two examples of features sets are shown below. Figure 5 shows the normalized octave spectrum from the bird chirp example in Figure 2. All spectrum are normalized such that the power contained in the spectrum is unity and is done to remove the effects of propagation distance on the recorded levels. Darker spectrum lines indicate that the data used for the that spectrum was from later in the file. Notice that there are no distinct features. Figure 6 shows the normalized octave spectrum from the aircraft example in Figure 3. Notice that the spectral shape is changing through the length of the time series data and that there is a distinct feature or hump around the 1 kHz octave band. For the aircraft example, the file contained 3,072,000 samples that were split into 1 second records overlapped by 25% of the record width for octave analysis resulting in 79 spectral estimates. The octave spectrum were computed for the octave bands between the 31.5 Hz band and the 16 kHz band resulting in 10 features per each spectrum. The spectrum shape from the two examples is visually different providing evidence that a machine learning algorithm should be able to separate the two feature sets as well.

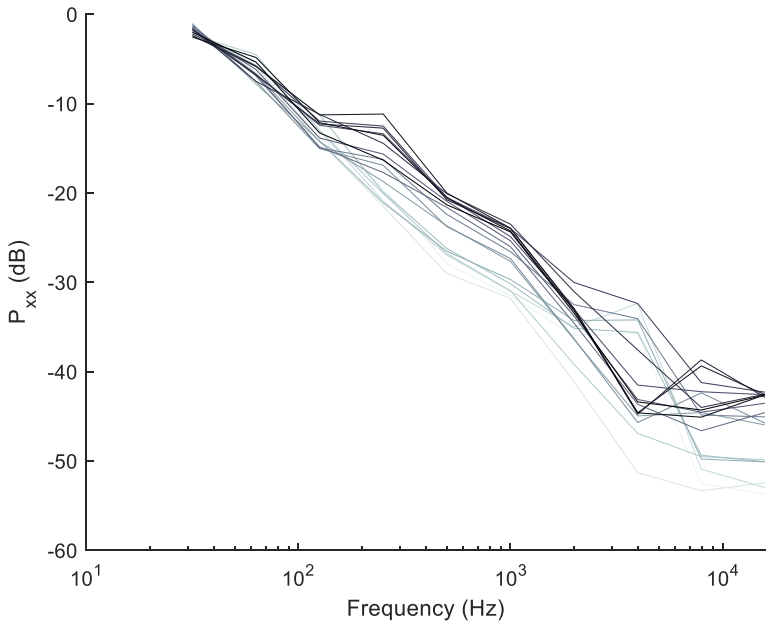

Figure 5: Octave spectrum of the bird chirp example (bird01.wav).

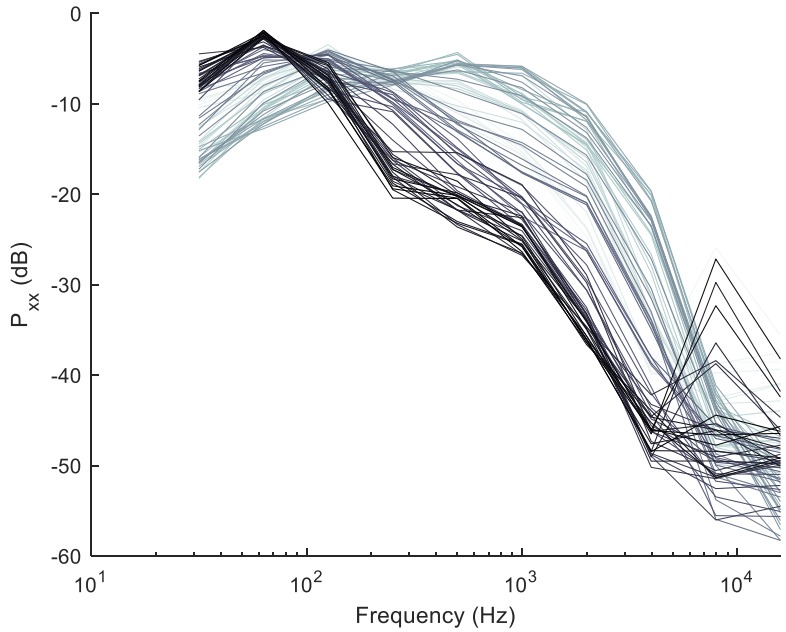

Figure 6: Octave spectrum of the aircraft example (plane08.wav).

The octave spectrum presented in the examples above demonstrate one of many different feature sets that could be used for classification of environmental noise. Other feature generation options include 1/n octave spectrum, which would break the octave bands seen above into smaller bands, cepstrum processing, short-time Fourier transforms, and wavelets. All the different feature sets have the potential to provide varying degrees of separation between the signals. This project should survey as many feature sets as practical to gain the understanding of which provide the best separation. Selection of the classification algorithm will also impact the system performance and should be studied as well. The various combinations of a feature set and classification algorithm will form a rich test matrix for this project to study.

Conclusions

The raw data needed for this project is represented as audio time series recordings and is readily available. The data set provided by the sponsor is representative but may need to be augmented to address the class imbalance. The data issues that are present are mostly confined to data management practices as numerous combinations of signals, processing techniques, and classifiers will need be tracked and the results reported on. With such a diverse set of features and algorithms to explore, this project presents a great learning opportunity for those interested in signal processing.