Use Data Budgets to Manage Large Acoustic Datasets

Introduction

Efforts to understand the health of the ocean have increased significantly in the recent past. These efforts involve among other things, using more sensors, and more sophisticated ones. Many institutes are installing ocean observing networks to collect long-term ocean data. Over the next few years we expect to see a massive increase in the quantity (and quality) of ocean data available for determining the ocean's health.

It is vital that we better understand the effects of human activities on the ocean, its flora and fauna. Recent legislation in Europe, has started a new wave of activity in this area.

Common to all these activities is the measurement of ocean sound, which is considered to be from three sources:

- *Geologic Sounds, produced by weather, seismic events, currents, etc.*
- *Biologic Sounds, produced by wildlife in the sea*
- *Anthropogenic Sounds, produced by human activity, such as shipping and construction*

The presence of anthropogenic sounds has triggered strong interest because of rising shipping, construction and offshore energy activities.

Light is absent in the sea, except near the surface, leaving sound as the key communication and navigation method for sea life. Many researchers are trying to understand how various species use sound, in order to determine safe anthropogenic sound levels.

The Problem

Hydrophones differ from many ocean instruments in that they produce dynamic data, sound pressures that vary at rates of hundreds of kilohertz. For example waveform data for high frequency clicking mammals, such as harbour porpoises, require sample rates of 512 kHz. This gives a data stream of 1.5 MB per second, or 130GB per day.

This is a massive quantity of data if the primary objective is to better understand the 'big' machine sound sources, which are typically less than 1000 Hz.

Many researchers are interested in sound that they can 'hear' using headphones. This sets another boundary for waveform data collection requirements.

Processing large datasets fully may takes years. Just managing and securing them is labour intensive. The unfortunate fact is that many wind up in a storage vault with no

post processing done to them. There may be important scientific information locked up in those vaults.

How do we ensure that all data receives at least some processing? How do we detect the anomalies in our datasets, such as an unexpected cetacean visit?

Smart Hydrophones

Typical analog hydrophones require the user to combine the necessary elements of a data collection system, including the pre-amplifier, filters, analog to digital converter, data acquisition software and storage. Combining all these elements and managing to collect and store trusted waveform data is a big accomplishment. System level calibration is often undertaken as an afterthought.

Digital hydrophones and dataloggers combine the analog and digital data path, giving us consistent acoustic data. In general, these instruments collect the waveform data well. Some dataloggers have a duty cycle function, where a number of minutes per hour are recorded, to reduce the quantity of stored data. Although statistically useful, the risk of missing important events, or cutting off those events is too high.

The Smart Hydrophone was developed to address data bandwidth challenges. The instrument combines an integrated digital instrument with acoustic processing capability and a large data memory.

A well designed Smart Hydrophone has

1. *High quality analog performance*
2. *An accurate, stable timebase that can be synchronized with other instruments*
3. *Low power, for autonomous operation, and performance over long cables.*
4. *A building block structure, allowing it to be integrated into larger systems*
5. *Signal processing and event detection*
6. *A robust data link to stream real-time data, and to retrieve large data records in a short time*

Acoustic processing in the instrument reduces the quantity of stored data while adding value to it.

Smart Hydrophones provide a toolbox that let users manage data collecting and post-processing needs through Data Budgeting.

Data Budgeting

Acoustic data budgeting requires that the project planner understand the data collection tools, as well as the required outcomes of the acoustic project. The outcomes may fall into two categories, immediate short-term requirements, and longer term loosely defined data for historical reference.

The immediate outcomes are typically linked to a specification or research target. They must be fully met with some check of the data. The longer term outcomes may be more useful for showing the context of the collected data.

Example

The data budgeting approach can be illustrated by an example. *In this one, the project requires that shipping traffic sounds be recorded, with post-processing showing the frequency of events greater than 100 dB re.uPa, and greater than 140 dB, at the measurement points. The required bandwidth for this monitoring is 800 Hz.*

A secondary requirement for this project is to determine if any sea mammals are present during shipping events. Of particular interest is the harbour porpoise. The maximum bandwidth is 180 kHz if the harbour porpoise is to be included in the study.

The project is to last for 30 days.

Solution A

This requirement suggests continuous recording of waveform data up to 180 kHz, with a sample rate of 512 kHz, giving 130 GB of data per day. For a 30 day project this totals 4 TB.

Processing of the 4 TB of data for vessel traffic requires many days, or weeks. Additional processing is required to look for sea mammal sounds.

Solution B

The following table shows some of the Smart Hydrophone tools available

Data Stream	Type	Logging
1	Store Waveform Data	*Continuous* *Duty Cycling* *Event Driven*
2	Store Spectral (FFT) Data	*Continuous* *Event Driven*
3	Store Event Results	*Event Driven*

Note that streams 1 and 2 collect data at different sample rates.

Since the vessel sound data is important, we acquire it continuously to ensure the 800 Hz bandwidth. Using a 2 kHz sample rate gives us maximum resolution sound data in this bandwidth.

The spectral data is still continuously collected at 512 kHz sample rate, stored four times per second, with a spectral bandwidth of 500 Hz (FFT bin size). This produces about 6 kB per second.

Total data size is 15 GB for the waveform data and 15.5 GB for the spectral data.

Post processing of the vessel sound data is simpler since it is stored at a much lower bandwidth. Processing of spectral data looking for sea mammal events is done using a math processing program or spreadsheet. Spectrograms can be charted directly.

Solution C

Combining the above solution with the Smart Hydrophone's event detection tools provides meta data that can enhance post processing, while reducing collected data further.

Assuming the vessel sound bandwidths are known (specified in the requirements), then the event detector triggering can be set up for them. The instrument logs waveform data when vessels are passing, eliminating 'dead' data from memory. Use duty cycling to ground truth the detectors, and ensure that some baseline data is collected. In this case collect 6 out of 60 minutes (10%).

Configure triggers for the harbour porpoise sounds as well. Continuous spectral data is collected as before, meeting that project requirement.

Data required for this setup is more complicated to predict, since it depends on vessel traffic. If we assume a busy ship channel where vessels are present for 40% of the time, waveform data memory totals 6.7 GB, and the spectral data remains at 15.5 GB.

Post processing now provides an Event History file that is imported into a spreadsheet or math analysis program to count and analyze the number of vessel passes and harbour porpoise activity. The file can be used to quickly locate important waveform data for detailed examination. This is performed in minutes rather than days. Validate the detectors by comparing with the continuous spectral data.

Summary

The memory used for the above solutions is gathered in the table below. Note that memory required for the Event History file is negligible.

More than two orders of magnitude separate the full bandwidth waveform approach, solution A, from the other solutions. This smaller data set size reduces post-processing effort as well.

Data	Solution A	Solution B	Solution C
WAV	4 TB	15 GB	6.7 GB
Spectral	-	15.5 GB	15.5 GB
Total	4 TB	0.03TB	0.02 TB

If the project calls for the use of multiple instruments, then the reduction in post-processing time is even more significant.

Conclusion

The introduction of large low cost data storage has led to many thinking that data memory is unlimited. This may be true, but it does not address the challenge of how to process and manage massive data sets.

The Smart Hydrophone tool box allows users, for the first time, to plan projects using Data Budgeting. This helps ensure all the important data is collected, while being manageable.

Producing meta data alongside the stored acoustic data speeds up the understanding of new data sets by providing an overview of the complete data record.