**Scientific Problem**

The goal is to extract information from sound recordings about animal community composition, biological and weather phenology, and natural and human disturbance impacts by analyzing soundscape recordings.

To this end, we use two types of measurements that quantify the level of acoustic activities:

- On a single recording using **alpha-diversity indices**
- Various alpha-indices measure different aspects
- E.g. Acoustic Complexity Index (ACI) measures the overall frequency variation in a spectrogram
- Between two different recordings using **beta-diversity index**

**Challenge: Big Data**

Recent advances in automated acoustic recording as shown in figures below, massive data storage, and rapid soundscape measurement result in considerable amounts of data.

- The Purdue Center for Global Soundscapes (CGS) library is composed of soundscapes collected at a global scale.
- Started in 2008, this data collection represents to date around 120 TB in wav format organized in 13 collections across hundreds of different sites around the world, and is still growing.

**Challenge: Sequential Workflow**

- The large number of collections of sound recording data are stored in a set of external drives.
- A powerful workstation is connected to the external drives.
- Scientists perform computation and analysis on the data using the workstation.

![Figure I: Traditional sequential acoustic analysis workflow.](image)

**Conclusions and Future Works**

Our modularized system enables flexible acoustic analyses workflows on HPC clusters.

- Our system allows the users to specify the type of analyses, specify what sound files of interest, automatically generate sets of PBS jobs, and submit PBS jobs that run the workflows in parallel.
- The ability to scale the workflows on HPC nodes allows acoustic analysis on big data that have never been tackled before at this scale, in a reasonable time frame.

Future works include:

- Integrate more workflows into our system;
- Integrate statistical and machine learning methods for future analyses.

**Acoustic Analyses Framework**

A modularized system that enables acoustic analysis workflows on HPC platforms over hundreds of terabytes of soundscape recording collections.

![Acoustic Analyses Framework](image)

**Case Study: Single Recording Acoustic Activity**

Alpha-diversity indices are calculated using our framework to measure the single recording acoustic activity. The following table presents 3 representative of the 11 alpha-diversity indices.

<table>
<thead>
<tr>
<th>Index Name</th>
<th>Calculation Principle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioacoustics indexBoeiman et al. 2007</td>
<td>Area under the frequency spectrum.</td>
</tr>
<tr>
<td>Temporal entropy (Ht)Sueur et al. 2008</td>
<td>Entropy of the amplitude envelope over the time units.</td>
</tr>
<tr>
<td>Spectral entropy (Ht)Sueur et al. 2008b</td>
<td>Entropy of the amplitude values over the frequency units. Spectrum maximized.</td>
</tr>
</tbody>
</table>

**Example**

- Pumilio database
- Acoustic Analysis Framework (AAF)
- PBS job
- Local Disk
- GPFS
- Worker
- Input File

**Figure II: HPC-enabled parallel acoustic analysis workflow.**

- Pumilio is a SQL database with metadata about the recording collections with web-based user interface.
- Our framework takes user specified options (e.g. locations and time periods), queries Pumilio, and submits jobs to a cluster.
- Each job runs R Workers that perform acoustic analysis in parallel.

**Results and Discussions**

Our framework is able to finish the workflow in a reasonable time frame as presented in Table II.

- The aggregated size of data is 44 TB.
- Our framework used 13 jobs to process 13 collections of recording data.
- Each job finished within 3 hours.

**Table II: Execution time of alpha-diversity indices calculation jobs.**

<table>
<thead>
<tr>
<th>Collection Name and Short Description</th>
<th>#Files</th>
<th>Size</th>
<th>#Nodes</th>
<th>Walltime</th>
</tr>
</thead>
<tbody>
<tr>
<td>arizona2013 - Long-term wildfire disturbance survey, desert ecosystems</td>
<td>121739</td>
<td>10 TB</td>
<td>24</td>
<td>2:18:05</td>
</tr>
<tr>
<td>Arizona - Short-term habitats description, desert ecosystems</td>
<td>531</td>
<td>46 GB</td>
<td>1</td>
<td>0:12:25</td>
</tr>
<tr>
<td>Borneo2014 - Short-term habitats description, tropical ecosystems</td>
<td>2663</td>
<td>252 GB</td>
<td>8</td>
<td>0:12:25</td>
</tr>
<tr>
<td>Chicago - Long-term habitats description, urban ecosystems</td>
<td>53947</td>
<td>1.5 TB</td>
<td>8</td>
<td>0:41:01</td>
</tr>
<tr>
<td>Hoosier - Short-term road disturbance survey, temperate ecosystems</td>
<td>15863</td>
<td>7.7 TB</td>
<td>8</td>
<td>0:07:45</td>
</tr>
<tr>
<td>Laselva2011 - Long-term habitats description, Tropical ecosystems</td>
<td>74204</td>
<td>3.7 TB</td>
<td>16</td>
<td>0:00:21</td>
</tr>
<tr>
<td>Leopold - Short-term road disturbance survey, Temperate ecosystems</td>
<td>22401</td>
<td>2.4 TB</td>
<td>8</td>
<td>1:06:53</td>
</tr>
<tr>
<td>tip2008 - Long-term habitats description, Temperate ecosystems</td>
<td>38464</td>
<td>5.0 TB</td>
<td>8</td>
<td>0:01:41</td>
</tr>
<tr>
<td>tip2009 - Long-term habitats description, Temperate ecosystems</td>
<td>32425</td>
<td>6.5 TB</td>
<td>8</td>
<td>2:23:39</td>
</tr>
<tr>
<td>tip2010 - Long-term habitats description, Temperate ecosystems</td>
<td>46937</td>
<td>5.8 TB</td>
<td>8</td>
<td>2:30:48</td>
</tr>
<tr>
<td>tip2011L - Long-term habitats description, Temperate ecosystems</td>
<td>47970</td>
<td>5.2 TB</td>
<td>8</td>
<td>2:32:14</td>
</tr>
<tr>
<td>tip2012 - Long-term habitats description, Temperate ecosystems</td>
<td>22475</td>
<td>2.5 TB</td>
<td>8</td>
<td>1:08:14</td>
</tr>
<tr>
<td>wells - Long-term habitats description, Estuaries ecosystems</td>
<td>41847</td>
<td>3.3 TB</td>
<td>8</td>
<td>2:05:40</td>
</tr>
</tbody>
</table>

**Acknowledgments:**

CNH-BCS NSF grant 1114945, AISL NSF grant 1323615, National Park Service, Purdue OCPRI’s incentive Grants Program, Department of Forestry and Natural Resources Wright Fund, College of Agriculture Information Technology, Research Computing, Information Technology at Purdue. The authors want to thank Dante Francomano and Jarrod Doucette for their valuable inputs and discussions. Code samples and timing data are available on github: https://github.com/purdue-research/RCAC-Staff/sound.record.