Prototyping and Production of a Source Finding Pipeline for GASKAP-OH

Principal Investigator: Jo Dawson

Proposal team:
Jo Dawson (Macquarie University/CSIRO), Shari Breen (SKAO), Beth Cappellazzo (Macquarie University/CSIRO), Jayender Kumar (CSIRO), Matt Whiting (CSIRO), Rina Kasai (Kagoshima University)

Commitment availability:
Combined 0.4 FTE from one PhD student (Beth Cappellazzo) and one postdoc (Jayender Kumar), plus 0.1 FTE from one survey PI (Jo Dawson). Note also that Matt Whiting is part of the `ASKAPsoft` team and will have high-level involvement throughout.

Problem Statement

The GASKAP-OH survey requires a reliable 3D source-finding pipeline that can be run efficiently on very large (2.2 TB) FITS cubes. The sources occupy a very small fraction of the total pixel count, have high dynamic range, and areas of the datacube suffer contamination from bright source sidelobes, which are picked up as numerous false detections and may obscure weak sources. The source-finding pipeline must address these issues, and produce source catalogues and cutout cubelets. 3D source finding is a post-processing step that is carried out on science-ready datacubes output from the standard `ASKAPsoft` imaging pipeline.

Project Overview

The Galactic ASKAP survey of OH (GASKAP-OH Dawson et al., 2024) is one of the nine approved ASKAP Survey Science Projects, and a precursor for future SKA science. The survey is observing the four 18-cm ground-state hyperfine lines of the hydroxyl radical (OH) in the Galactic Plane, Galactic Bulge, Galactic Centre and the Large Magellanic cloud. We target two classes of object: bright, compact OH masers and extended OH absorption against background radio continuum sources. OH masers probe high-mass star formation, evolved stellar mass loss, supernova remnants and the structure of our Milky Way Galaxy; while OH absorption probes the physics and evolution of the molecular interstellar medium (ISM). Together with our sister survey, GASKAP-HI (Dickey et al., 2013), we link the evolution of gas from the warm atomic medium, to cold atomic gas, to molecular clouds, star formation, and finally mass-return to the ISM.

As the highest-frequency (1.6–1.7 GHz), highest spatial resolution (10 arcsec) and highest spectral resolution (0.579 kHz = 0.1 km s^{-1}) survey with ASKAP, and covering the complex emission of the Galactic Plane in both Stokes–I and Stokes–V, we are one of the most technically challenging ASKAP surveys. This means GASKAP-OH is a critical testing ground for SKA data processing and science analysis. Standard `ASKAPsoft` processes are used for imaging but have proved unsuitable for 3D maser source finding. Our data cubes are very large (>2 TB), and contain complex structures on many scales (that are challenging to image), as well as extremely bright sources that contaminate our images and cubes with sidelobes. These factors all complicate source finding, both from an algorithmic perspective, and from a resourcing/compute perspective. We note that our needs are quite distinct from WALLABY, whose sources are weaker and more extended in both space and velocity, and do not suffer the same level of sidelobe contamination as the bright masers in GASKAP-OH.

GASKAP-OH will run two classes of source finding: one for OH absorption and one for OH masers. **Only maser source finding is the subject of this EoI**. For OH absorption, it is sufficient to identify regions of bright continuum emission in 2D images (since these are the only places where absorption is detected). This is essentially solved: the standard output of the `ASKAPsoft` `Selavy` (Whiting & Humphreys, 2012) continuum source finder is sufficient, and this is run as part of our standard imaging pipeline. In contrast, maser source finding must be carried out on the 3D spatio-velocity cubes by identifying islands of contiguous emission. While this is not a new problem, there are some complexities. Masers are point sources that occupy a very small fraction of the total pixel count, but they can be bright enough to contaminate significant regions of the datacube with extended source sidelobes, which are picked up as numerous false detections. Our large cubes also complicate implementation. Initial testing on setonix with `Selavy` has highlighted two main issues: (a) The standard `Selavy` implementation is unable to reliably reject false detections without compromising our ability to pick up weak sources, (b) our large cubes are problematic: `Selavy` has so far been unable to complete processing of a full cube on setonix before jobs time out, despite parallelisation.

Here we request AusSRC support in three main stages:

- **Science Support:** determining an appropriate source-finding algorithm and input parameters, including development and testing of a sidelobe rejection protocol.

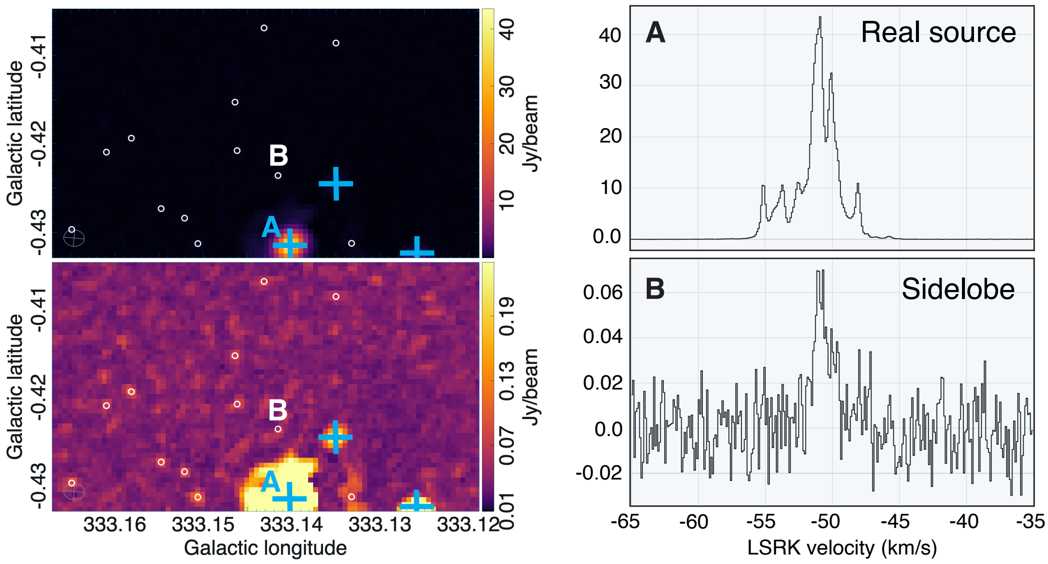

Figure 1: Example of `Selavy` 3D maser source finding results from a single test run. The two left panels show zoomed-in peak flux density images of a small portion of a GASKAP-OH subcube. (The size of these panels is 0.003% the area of a single GASKAP-OH footprint.) The colour-scale on the top panel is adjusted to the brightest source; the bottom panel is saturated to show weaker sources and sidelobes. The blue crosses show real detections and the white circles are sidelobes (false detections). The panels on the right show example spectra from the brightest source (A) and a sidelobe (B). Note that we expect real sources in our cubes as weak as this sidelobe.

- **Software development:** Ensuring the approach is scalable and can be parallelised, and can work efficiently on the full 2.2 TB cubes.
- **Workflow/pipeline development:** Stitching together source-finding, catalogue creation and cubelet creation into a production pipeline.

Existing resources or projects

GASKAP-OH team members have access to Setonix compute resources under the *ja3* project. This is a joint project providing suppport for multiple teams – GASKAP-OH's nominal share of this is 1 MSU (of the 13.7 MSU for 2025). This project is subject to yearly application (led through CSIRO by Lead CI Whiting), and so carries the risk associated with each application. The *ja3* project provides access to the /scratch filesystem and acacia for workspace and storage, as well as the `ASKAPsoft` processing package (providing `Selavy` - note that one of the team members, Whiting, is the lead `Selavy` developer.)

In this environment, we have run initial 3D source-finding tests with `Selavy`, using our initial commissioning observations. The cubes being searched are large – 2.2 TB – and so a challenge has been how to parallelise the work efficiently so that the source-finding job completes within the permitted time. `Selavy` allows for parallel access, but finding the appropriate settings that balance compute and I/O is still a work-in-progress. Our tests to date have been on small subsets of the full cube. We have also trialed a sidelobe rejection method based on cross-correlating the spectra of identified "sources", which works on the basis that sidelobes should share spectral characteristics with their originating source. Initial results are promising, but development is in its nascent stages and the approach needs dedicated testing and refining.

We are also in discussion with the WALLABY team on their `SoFiA` source finding algorithm. Quantitative assessment of the relative merits of `SoFiA` and `Selavy` would form part of this AusSRC project. Note that the GASKAP-OH requirements are somewhat different to those of WALLABY, in that the sources are compact but often very bright, in contrast to the fainter, more extended nature of HI galaxies.

Extra information

The GASKAP-OH team are collaborating on another AusSRC EoI with WALLABY and the CSIRO CASDA team to implement improved cubelet cutout functionality in CASDA.

References

Dawson, Breen, & the GASKAP-OH team. 2024, in IAU Symposium, Vol. 380, Cosmic Masers: Proper Motion Toward the Next-Generation Large Projects, ed. Hirota, Imai, Menten, & Pihlström, 486–490

Dickey, McClure-Griffiths, & Gibson et al. 2013, PASA, 30, e003

Whiting & Humphreys. 2012, PASA, 29, 371