

A Human-In-The-Loop Experiment to Investigate the Effect on Detection Performance of Having Simultaneous Access to Multiple Sonar Sensor Systems in a Single Display

Rebecca Kuster

Maritime Operations Division,
DSTO, Edinburgh
PO Box 1500, Edinburgh SA
5111

rebecca.kuster@defence.gov.au

Andrew Knight

Maritime Operations Division,
DSTO, Edinburgh
PO Box 1500, Edinburgh SA
5111

drew.knight@defence.gov.au

Ben Fletcher

Maritime Operations Division,
DSTO, Edinburgh
PO Box 1500, Edinburgh SA
5111

ben.fletcher@defence.gov.au

Abstract

We report results of a human-in-the-loop experiment that aimed to quantify the effects of networking diverse sonar sensors in an operational environment. Historically, sonar performance has been evaluated by conducting a series of detection runs to measure the range at which a fully alerted operator can detect a contact using the sensor being tested. Two effects precluded such a simple measurement in our case: (1) fully alerted detection ranges can vary considerably from those achieved by unalerted operators, and (2) we needed to evaluate the effect of operators having access to information from more than one sensor. Operator performance was measured by replaying sonar data recorded at sea and displaying the resulting information on a single operator's display. Participants, consisting of DSTO staff, were given standardised training in the operation of the display, and were asked to identify contacts of interest. We investigated a number of aspects of their performance as the number of sensors available on the display was varied. The key performance metric was the time participants took to identify contacts of interest. From this small-scale experiment (single operators before a single screen), we identified issues associated with sensor networking, training, situation awareness, information fusion and information overload. Quantitative and qualitative results derived from the experiment indicate the challenges of conducting experiments with a limited number of participants and recorded real-world data.

1 Introduction

The experiment described in this report was conducted to support a DSTO-developed Capability Technology Demonstrator (CTD) on "network-enabled undersea warfare". This CTD, which networked a diverse range of underwater sensors, aimed to improve the capability of ADF surface platforms to detect underwater contacts of interest (such as submarines). There was also potential for a reduced crewing requirement if information from several sonar displays (usually crewed by multiple sonar operators) could be integrated into one display. This experiment was conducted to assess the Sonar Test Bed (STB), to examine the effects of sensor networking and to highlight areas where future work would be beneficial.

The primary issue to be investigated was the benefit of connecting several undersea sensor systems to a single operator's display, and, in particular, whether this improved detection performance in an operational setting. In other words, our experimental question was: *Does networking a number of sonar systems facilitate more accurate and timely assessment of what is and isn't a contact?* This aim required human operators to decide if contacts that appeared on the screen were to be classed as "contacts of interest".

2 Experiment

This experiment was based around feeding data recorded during a sea trial into the STB. The data was able to be re-played as required by the experiment; that is, the data could be re-played by switching on the data stream from 1, 2, 3 or 4 sonar sensors in any combination. The combination of sensors available was limited to those that were recorded and which were detecting the target(s) at the time during the sea trial. The sensor networking used in the trial (that is applicable to the experiment) can be seen in Figure 1 below. Data from all 4 sonar systems was relayed to the ship, where it was recorded. During the experiment, participants played the role of sonar operators on-board HMAS Arunta.

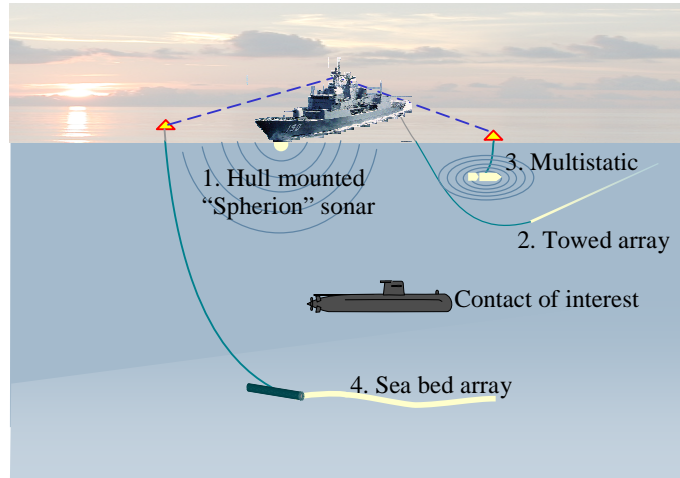


Figure 1: The network of sensors used in the sea trial.

The sonar systems used in the trial, and replayed during the experiment were:

1. a hull-mounted (Spherion) sonar, mounted on the bow, as shown in Figure 1,
2. a towed array (Advanced Surface Ship Towed Array Sonar System or ASSTASS)—a long flexible hose containing an array of hydrophones, and towed through the water by the ship,
3. a multi-static sonar system, in which the sound source and the receiver are physically separated, and
4. a sea bed array—DSTO’s Deployable Remote Monitored Array (DRMA)—deployed on the sea bed.

The Spherion and towed array (ASSTASS) were used in both active and passive modes during the experiment. Multi-static was an active source, with the Spherion, DRMA, and towed array receivers detecting the noise in which ever mode they were in. The DRMA was passive-only.

The STB used in the experiment had a single display divided into various parts as can be seen in Figure 2 below.

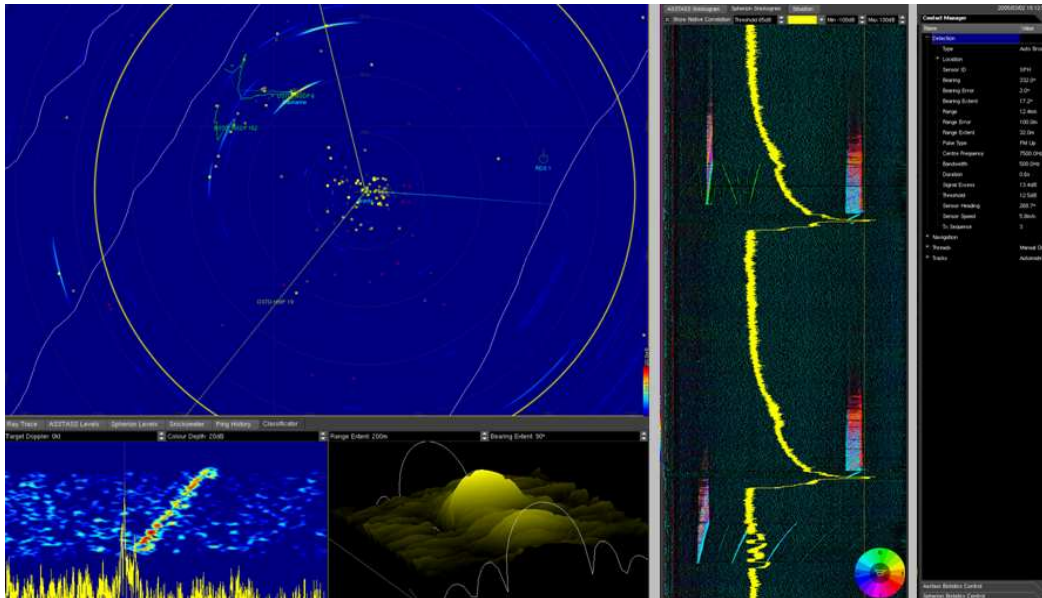


Figure 2: Screen capture from the STB.

The main part of the display, the Plan Position Indicator (PPI), is a plan (top) view of the area of interest. In Figure 2 above, it is the top-left, largest, part of the screen. It brings together information from the various sensors being used into a single display. Of primary interest are the automatically generated detections, which

are colour coded according to the sensor that detected them. Participants were required to investigate these automatically generated detections and using the display tools, come up with an assessment (according to a 3-point confidence scale¹) of whether they were contacts of interest.

2.1 Experiment Design

The experiment was designed around 6 data sets that were chosen because they contained good detections on the contacts of interest on multiple sensors. The number of sensors in the data sets was varied.

The experimental sequence was planned so that participants never worked with the same at-sea data set twice (say, with different configurations of sensors). This was to avoid an expected learning effect. Eleven participants were obtained from DSTO staff, which included two embedded naval personnel designated as ‘experienced participants’.

A standardised training procedure was put in place for all participants. Participants were also asked to complete a questionnaire on the useability of the system. A photograph of the experiment being conducted can be seen in Figure 3 below.

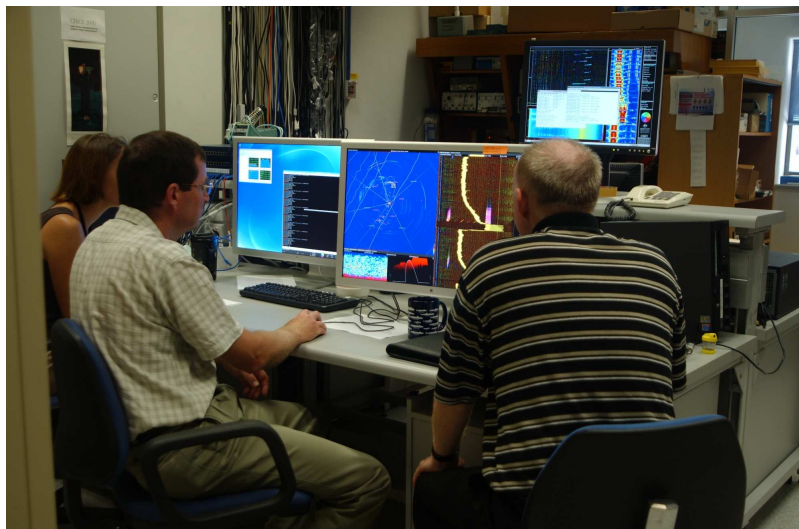


Figure 3: A photograph of the experiment being conducted. The participant in the centre is investigating detections to determine contacts of interest while the observers on either side record information.

To balance the need for multiple replications with a single participant and the need to minimise participant fatigue, participants were asked to do four experimental data sets each. It was estimated that these and the practice data sets would take 1½ hours (6 data sets in one session). So, the experiment was designed around four runs per participant using all six of the data sets, as we wanted to ensure that the experiment covered a variety of conditions. The expectation was that the variation in data set and participant should be random or able to be extracted, allowing the effect of the networking to be analysed.

The order in which the experimental data sets were undertaken was also varied, along with the order of the different sensor-network configurations. This was to ensure that any effect due to order was minimised. Four sensor network configurations were used (1) ASSTASS (towed array) only, (2) Spherion only, (3) Spherion plus ASSTASS (towed array), and (4) all available sensors (which could be 2, 3, or 4 sensors, depending on the data set). A schematic diagram showing the replay of the recorded data using the STB, and the adjustment of sensor type is shown in Figure 4.

¹ I.e. “Not confident”, “Maybe”, or “Yes, confident”.

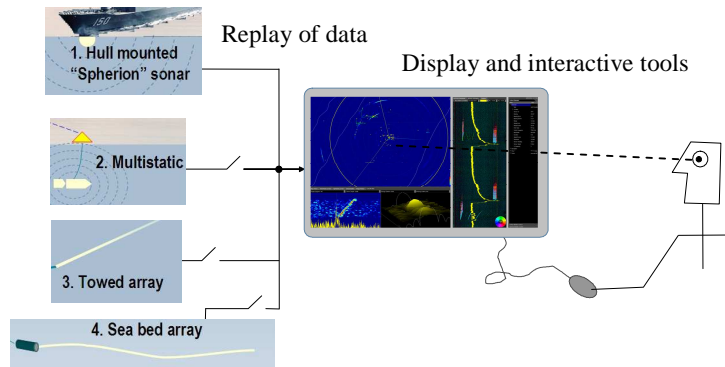


Figure 4: Schematic diagram of the replay of sea-trial data to experimental participants using the STB and varying the available sensors.

It was important to do the Spherion plus towed array case (configuration 3) along with each single sonar system separately (configurations 1 and 2), because this allowed all contributions from ASSTASS and Spherion to be accounted for without any confounding sensor factors (although there is still the variation in participant and data set to take into account). The fully networked case was still run as part of the experiment, because it was considered important to test the concept behind the STB.

2.2 Measures of performance

The primary question was whether networking the surface sonar systems being investigated in this experiment improve operators' detection performance, in particular by providing more timely and accurate assessment of what is and isn't a contact. It was proposed that performance was primarily related to the time taken to detect underwater contacts². So, the primary measure of performance was the *identification time* (or ID time), equal to $t_2 - t_0$, where t_2 was the time at which the participant assessed a contact as being of interest (by declaring "Yes, confident" according to the 3-point confidence scale described above), and t_0 was the first time the contact appeared to fully alerted operators on that particular sensor. The quantity t_0 was determined by the trials team after close examination of the trials data. It was the earliest that they could possibly confirm that a detection recorded by a particular sensor corresponded to the contact, being fully alerted to look in the known location of the contact. The identification time would be used to compare the reporting delays of single sensor runs and multi-sensor runs.

The secondary measure of time used was the *confidence time*, which was the time the participant took to become confident that a contact was of interest after first taking an interest in it. The confidence time was equal to $t_2 - t_1$, where t_2 was the time at which the participant assessed a contact as being of interest (by declaring "Yes, confident"), and t_1 was the time at which a participant first took an interest in the contact. The relationship between these time delays and the observed events is shown in Figure 5.

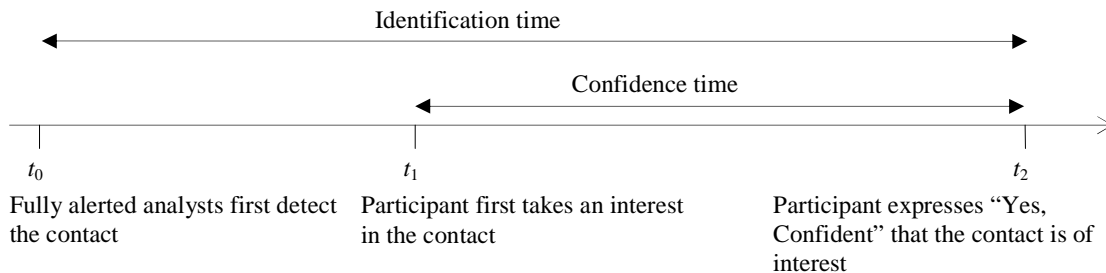


Figure 5: Definitions and relationships between key event times in the experiment.

² Traditionally, range of detection is the primary metric in sonar assessments. However, the relative positioning of the platforms during the trial precluded using range for this experiment.

The proportion of correct and false detections are important indicators of detection performance. The maximum possible number of contacts varied in the different data sets. The proportion of correct detections was therefore defined as:

$$\text{Correct detections (indexed)} = \frac{\text{number of correct detections}}{\text{maximum number of correct detections for that data set}}.$$

For instance, if there were two contacts able to be detected for the data set, and only one of these was detected, then the Correct detections (indexed) would be equal to 0.5.

A false detection was when a participant assessed that they were confident that a detection was a contact of interest, however, the detection did not correspond to a real known contact when compared to the truth data. The number of false detections could not be indexed in the same way as the correct detections, however, it was sometimes used in comparison with the number of correct detections to see whether the number of false detections was related to the number of correct detections, for example would having twice as many correct detections be at the cost of twice as many false detections? This metric was defined as follows,

$$\text{Correctness Ratio} = \frac{\text{number of correct detections}}{\text{number of false detections}}.$$

For each participant, the number of automatically generated detections they investigated was recorded, as well as how many of these the participant assessed as *No*, *Maybe*, or *Yes, Confident*. The number of *Yeses* compared to the total investigated might indicate participant confidence.

3 Results

3.1 Data set characteristics

Each of the data sets were recorded under slightly different conditions (for example, the acoustic environment varied) during the trial. The differences between the data sets, both quantitative and qualitative have been captured here, as they are an important factor affecting the results.

Some of the key differences between the data sets can be seen by examining the participant results for certain metrics. The following box plots show the variation of the identification time, and the confidence time grouped by data set. The box plots show the minimum and maximum values at the extreme ranges, the median observation, and the 1st and 3rd quartile observations. The individual observations (raw data) are also plotted.

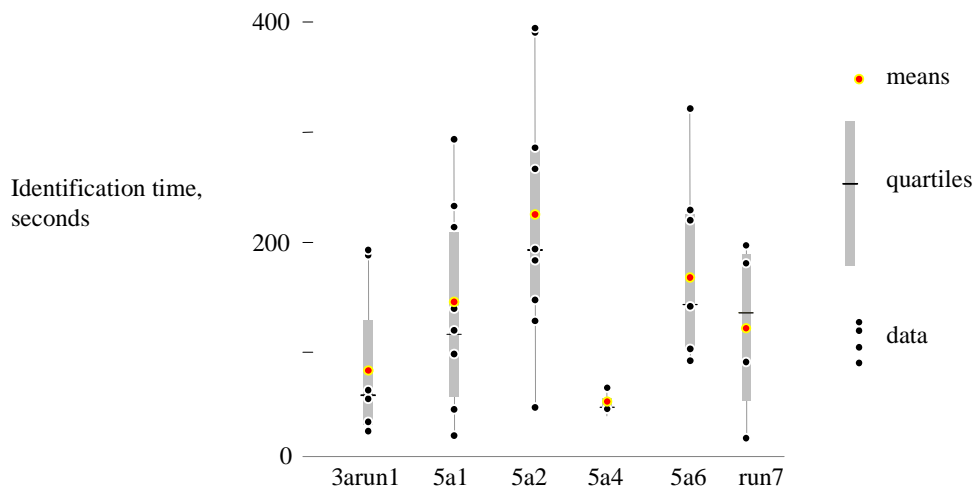


Figure 6: Box plot of identification times grouped by data set with statistical summaries.

Figure 6 above shows the variation in the time participants took to identify contacts of interest in the data sets. It can be seen that data set 5a4 stands out in comparison with the others as having very little variation in the time taken between participants. In contrast, data set 5a2 had a high level of variation, with observations at very short and very long time intervals. Figure 6 clearly shows that the variation of the time taken to identify contacts of interest varied significantly between data sets.

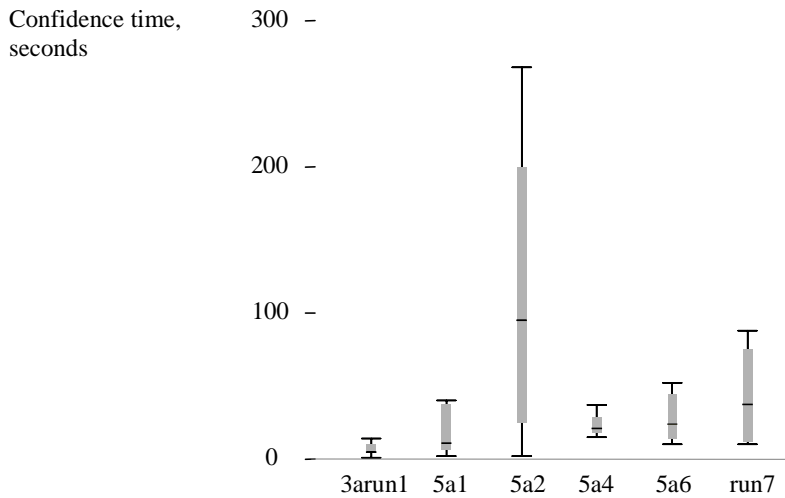


Figure 7: Box plot of confidence times grouped by data set.

The box plot in Figure 7 also shows variation in the time to gain confidence across data sets. The exception is data set 5a2, which had a large variation in the time participants took to gain confidence. This is consistent with the large variation in the identification time.

Figure 8 below compares the correctness ratio (O) (i.e. the ratio of correct (➡) to false (➡) detections) across data sets.

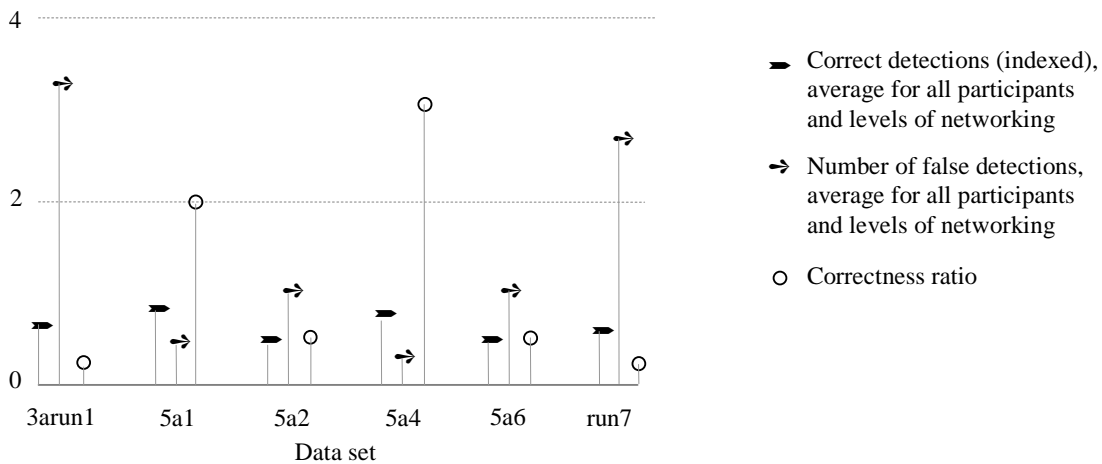


Figure 8 Comparison of the data sets looking at correct (indexed) and false detection, as well as the ratio of correct to false detections.

The number of false detections varies across data sets, with run7 and 3arun1 in particular having high numbers of false detections on average. The other characteristic that is highlighted in this figure is the high correctness ratios for data sets 5a1 and 5a4. Some runs had a significantly higher (or lower) number of automatically generated detections shown on the screen compared to others, which may have made it more (or less) difficult for participants to choose potential contacts of interest. Is there any evidence for this?

To answer this question, we note that runs 7 and 5a6 had relatively *more* detections on the display, yet in Figure 8 we see that these two data sets only constitute two of the four sets having the lowest correctness ratios. Conversely, runs 5a1 and 5a4 had relatively *few* detections on the display, and indeed (Figure 8) these two data sets have the two highest correctness ratios. So the evidence is positive that sparse displays lead to high correctness ratios, but not that highly cluttered displays lead to low correctness ratios. Low correctness ratios may be caused by factors other than display clutter.

3.2 Participant Characteristics

The participants were profiled against various metrics to identify any outliers or trends that would affect results. Figure 9 below shows the participants' identification and confidence times averaged over data sets and levels of networking. The resulting variance between participants of these two measures is not statistically significant. Both of the experienced participants took less time than the inexperienced participants to identify contacts of interest and to gain confidence. For each participant, the gap between the identification time and the confidence represents the time between the earliest possible detection time and the time the participant first takes notice of a detection (see Figure 5). This quantity is also smaller for the experienced participants.

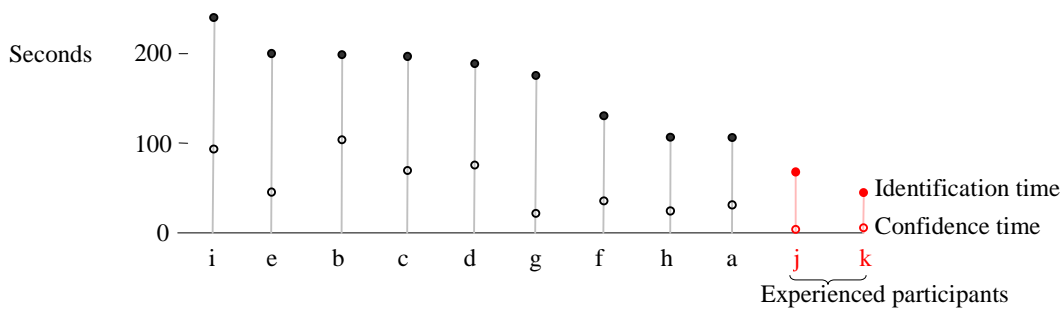


Figure 9: Identification and confidence times for participants averaged over all data sets and levels of networking, and plotted in rank order.

Figure 10 below profiles two metrics that are participant-dependent, so the variation between means of these metrics is statistically significant. The two metrics are the number of contacts investigated and the number of false detections (i.e. the number of false *Yeses*). Participant 'f' had an unusually high number of false detections. Participant 'i' investigated very many detections while having the longest mean identification time (Figure 9).

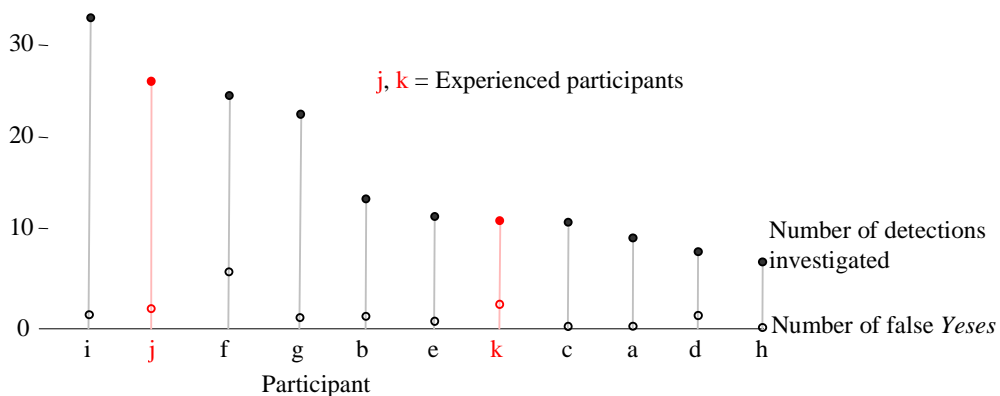


Figure 10: Number of detections investigated and the number of false *Yeses* for each participant, averaged over data sets and levels of networking, and plotted in rank order of the number of detections investigated.

3.3 Timeliness and Accuracy Results

The primary aim of the experiment was to address the question: *Does networking a number of sonar systems facilitate more accurate and timely assessment of what is and isn't a contact?* Regarding the timeliness aspect, the following figures show the identification times and confidence times plotted against the level of networking.

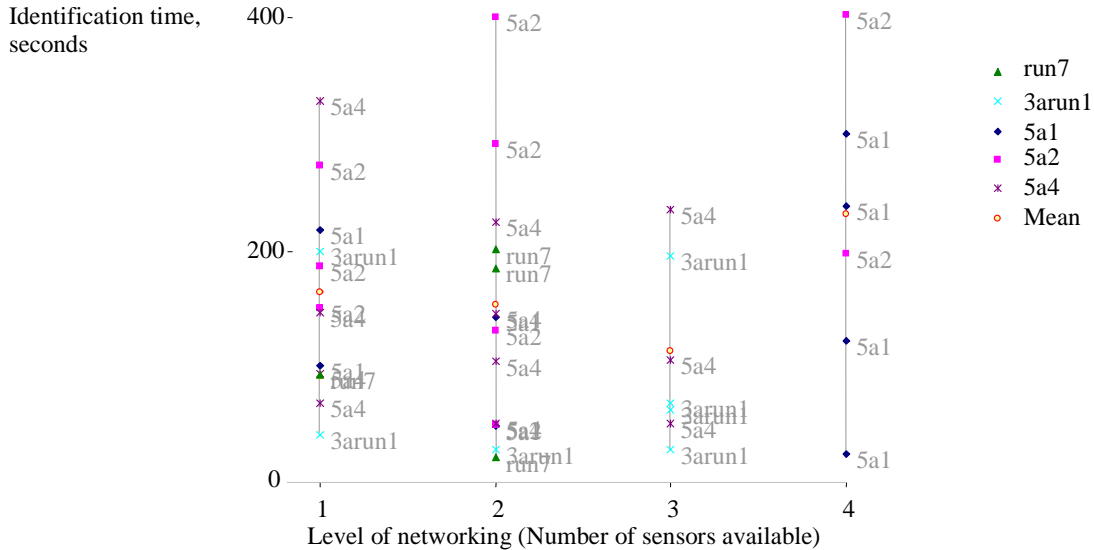


Figure 11: Identification times for each participant and data set plotted against level of networking.

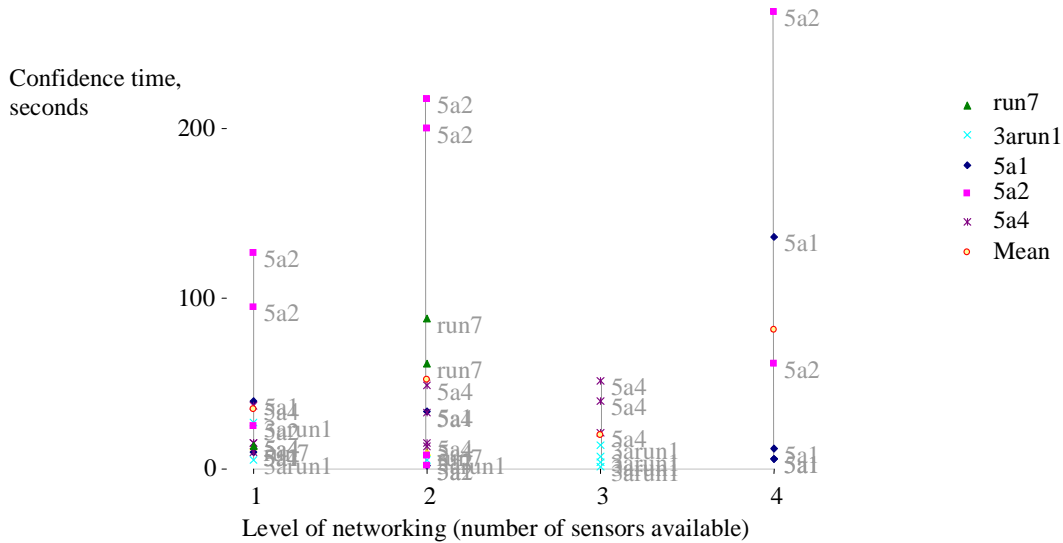


Figure 12: Confidence times for each participant and data set plotted against level of networking.

Although the mean times in Figure 11 might seem to indicate a decrease in identification time as the level of networking increases from 1 to 3 sensors, followed by an increase in time from 3 to 4 sensors, the underlying variance in the data prevents us from concluding that the level of networking either increases or decreases this time delay. Similarly, the variation in means in Figure 12 is not statistically significant.

From the last two figures we can see an increase in the time metrics as we go from 3 sensors to 4. Some participants expressed the opinion that the data sets were more difficult with 4 sensors present, so it may be

that the complexity of the information presented was a contributing factor to this increase in time metrics. It could also be due to the characteristics of the particular data sets that had four sensors. However, the difference is not statistically significant, and it is not possible to determine the cause of the increase without further data gathering.

Having dealt with the *timeliness* of determining a contact of interest, we now go on to investigate associated *accuracy* metrics, that is, the numbers of correct and false *Yes* decisions (here we will call a “*Yes, Confident* declaration” a *detection*).

Figure 13 shows a large increase in the mean number of contacts correctly identified when going from one sensor to two. When only one sensor was available (level of networking equal to 1), there were 11 occasions when participants failed to make any correct detections. At higher levels of networking, at least one of the two detectable vessels was always detected. This largely accounts for the increase in the mean number of correct detections when going from a level of networking of 1 to 2 or more.

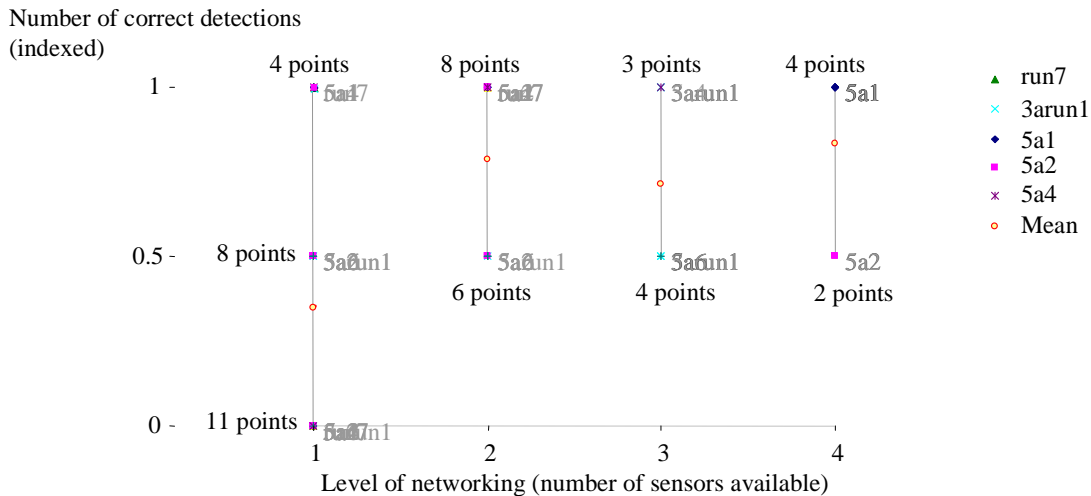


Figure 13: Number of correct detections for each participant and data set plotted against level of networking.

This trend is supported by the analysis of variance and other analysis conducted, which shows that the difference between one sensor and 2, 3, or 4 sensors is statistically significant (the differences in means between 2, 3 and 4 sensors is not statistically significant).

Figure 14 shows the (participant-dependent) number of false detections plotted against the level of networking. The differences in means for these measurements is statistically significant. However, this metric is participant-dependent, and furthermore, a single data point at a level of networking of 3 appears to be non-typical (8 false detections for one participant in dataset “3arun1”, whereas no other participant made more than 3 false detections and the mean for these other participants for that data set was 2—with no average for other datasets exceeding 1.6). This non-typical datum grossly skews the mean for a level of networking of 3. It is therefore likely that the statistical significance found for the variation in these means is caused by the non-typical result, along with other participant-characteristics, and is not a result of the level of networking.

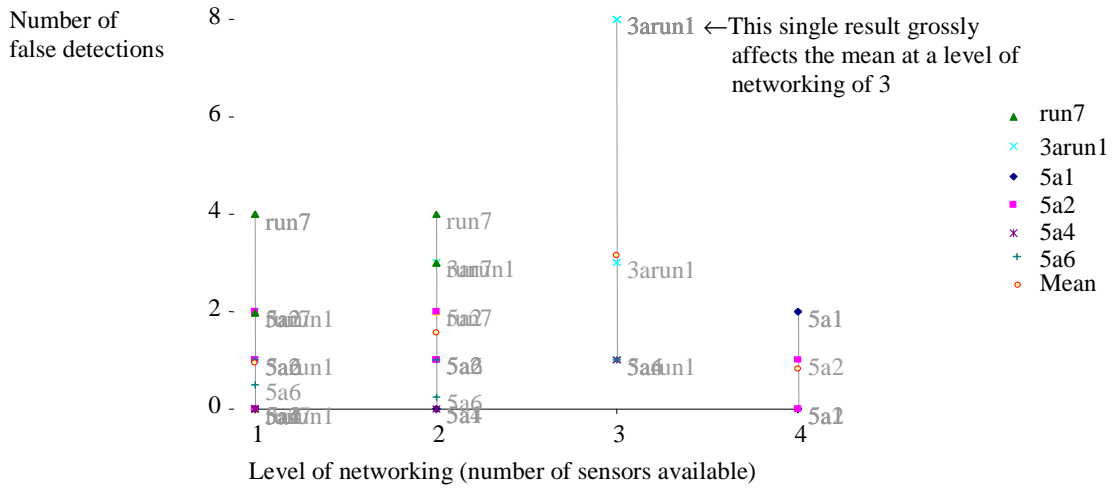


Figure 14: Graph of the number of false detections for each participant and data set plotted against level of networking.

Figure 15 shows a face-plot of the identification times for each participant and dataset against the level of networking. We can see (as noted in the discussion of Figure 13) that all detection failures occurred with only one sensor, indicating that there is a benefit of having more than one sensor present. This benefit may be a result of different sensors picking up contacts using diverse sets of noise sensitivities, directional discrimination, frequency sensitivities, etc.

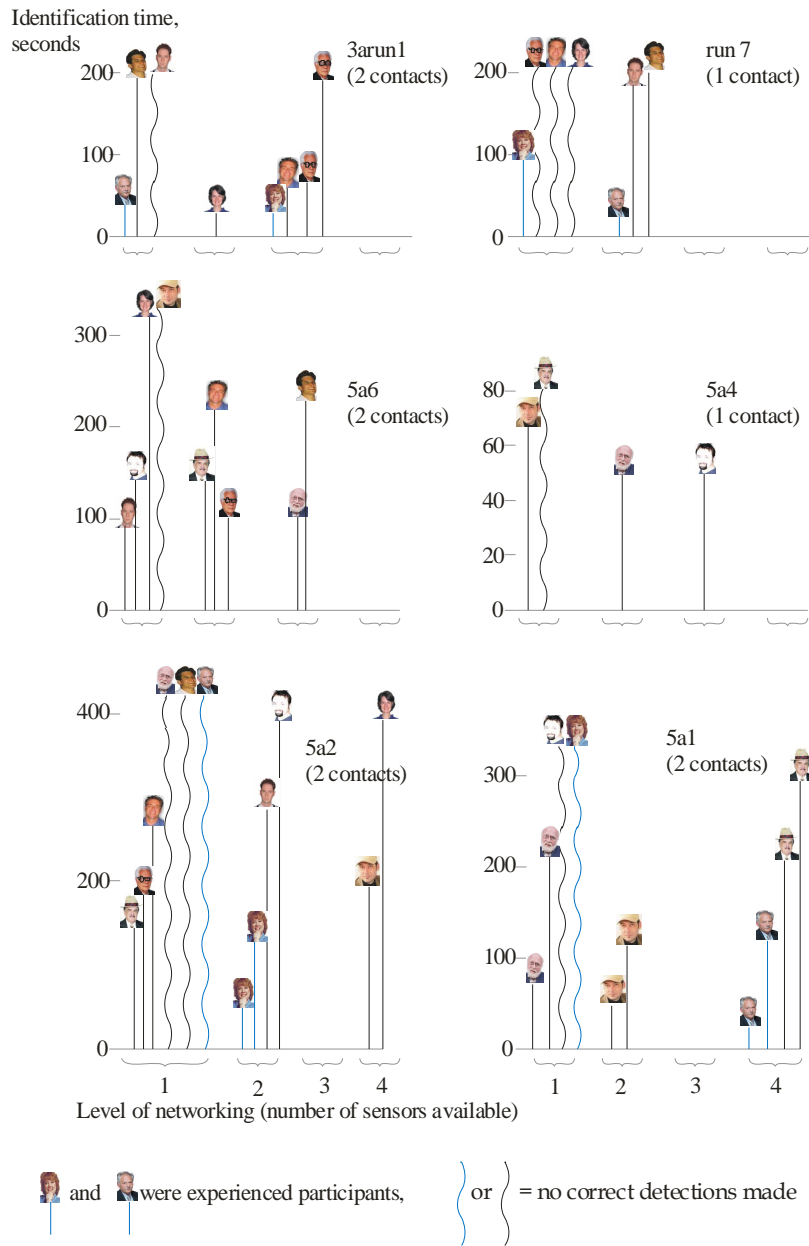


Figure 15: Identification times plotted for each data set, with the participants labelled by anonymized faces. Wavy lines indicate a failure to make any correct detections. (Note the variation in identification time between the data sets.) Duplication of a participant appearing in the result from a single data set indicates that there were two contacts that it was possible to detect.

It is of interest to examine the effect of level of networking on participants' confidence in declaring contacts of interest. As an indication of the participant's level of confidence, we took, the number of Yeses that a participant declared as a proportion of the total number of contacts they investigated. Level of confidence alone is of no particular benefit unless it is associated with a high rate of accuracy. We take the number of correct detections as a proportion of the total number of contacts investigated as an indication of a participant's rate of accuracy. Figure 16 shows a plot of levels of confidence and rates of accuracy plotted against level of networking.

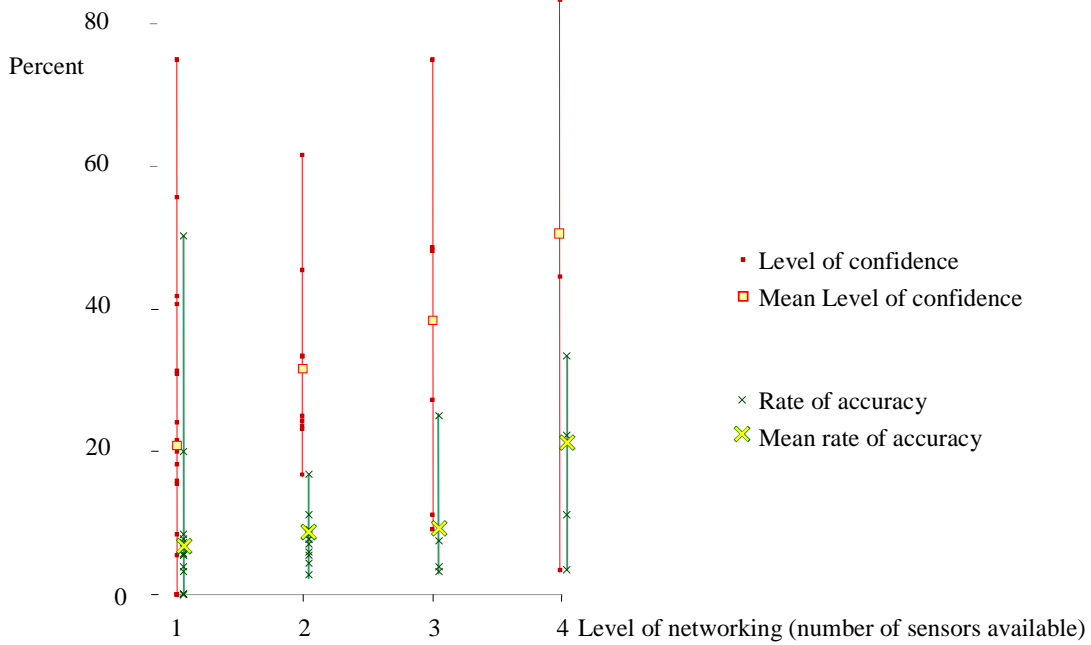


Figure 16: Participants' level of confidence and rate of accuracy plotted against level of networking.

We see that there is a trend of increasing mean level of confidence with increasing level of networking. There is a large variance in the underlying data, but analysis of variance (ANOVA) analysis³ shows a statistically significant difference in the means between levels of networking of 1 and 4 sensors. The mean rate of accuracy also increases with level of networking, however, this visual trend is not statistically significant, and more data would be required to confirm this result.

The plot in Figure 17 shows the correlation between these two measures, which is statistically significant.

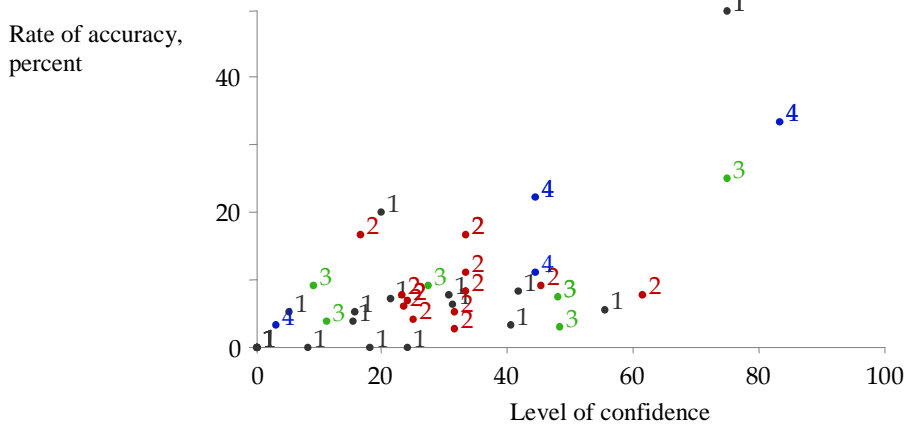


Figure 17: Rate of accuracy plotted against level of confidence, labelled by level of networking.

Five of the 11 participants (45%) commented that they felt they didn't have enough time during the experiment to further use various tools. However, while cognitive load was assessed to be between moderate and high on average by participants, some participants did not feel rushed or pressed for time as they examined less detections in a more methodical manner. Of note, the experienced participants were not among this group. Table 1 shows that the participants who felt pressed for time actually had a higher number of false detections and a lower number of correct detections on average.

³ A common technique used to determine statistical significance.

Table 1: Table showing correct and false detections by groups of participants determined by their perceived time-pressure.

	Average correct detections	Average false detections
Those who commented they felt pressed for time:	0.93	2.05
Those who did not comment on time constraints:	1.05	0.84

4 Conclusions

This experiment aimed to answer the question:

Does networking a number of sonar systems facilitate more accurate and timely assessment of what is and isn't a contact?

We cannot conclude from this experiment (given the variance of the limited datasets and of participant performance) that the level of networking either decreases or increases the timeliness of detecting contacts of interest. There is, however, evidence that this kind of networking increased the accuracy of participants' declaration of what was and wasn't a contact of interest. This benefit is clearly supported as the number of sensors increases from one to two, and is also supported by qualitative comments from participants stating that there was a definite benefit for them having two sensors rather than just one. When participants only had access to one sensor, they frequently failed to make any correct detection at all, but when they had access to two or more sensors, they never failed to make a correct detection. The results also indicate that networking more sensors gives less false detections.

Qualitatively, most participants coped well with the complexity of the display and the information provided. Feedback from the questionnaires indicated that, overall, participants felt challenged cognitively, but the display was easy to use, and they understood the tools and displays well. Complexity was observed to be of more concern when there was a high degree of networking present, and was also very data set dependent. Conversely, for some data sets at a low level of networking, participants were very comfortable. From an experimental viewpoint, this is a positive outcome, as the participants experienced a range of conditions similar to a real situation.

The use of human-in-the-loop (HITL) experimentation in this CTD was valuable for gaining feedback from the participants using the system. In this case there were many constraints, some of the key ones caused by the need to avoid the learning effect of presenting participants with the same data set.

Further work needs to be conducted to fully answer the questions proposed. Nevertheless, the experiment was valuable as an initial study to identify some of the issues and potential ways to overcome or avoid them. Qualitatively, the system shows great potential for future benefits to the Navy in the detection of contacts; and allows us to investigate reduced crewing and space requirements; and integration or fusion of information to provide a more comprehensive underwater picture leading to improved situational awareness. This experiment was exploratory only and not extensive enough to rigorously investigate these potential benefits.

Finally, although using recorded sea-trial data adds credibility to this kind of exploration, it comes at the cost of a loss of control of the configuration of sensors and contacts, and an inability to perform replications. It could be that using simulated data would allow a more conclusive experiment of this type.