VITECH INNOVATIVE
RESEARCH AND CONSULTING
15-9080 PARKSVILLE DR, RICHMOND, B.C.,CANADA V7E 4N9
TEL:1-604-241-5810
EMAIL:VITECH@APEXLINK.CA

PSC 2007 SOUTHERN FUND
PROJECT CLOSURE REPORT

A Feasibility Study of Using DIDSON Imaging Sonar to Estimate Species Composition at Mission

PREPARED FOR
PACIFIC SALMON COMMISSION

APRIL, 2008

TABLE OF CONTENTS

1 ABSTRACT

The hydro-acoustics program operated by the Pacific Salmon Commission provides estimates of the daily passage of salmon at a site located near Mission B.C. The focus for Fraser River Panel management is sockeye and pink salmon and the estimates of daily passage for these key species are determined through an analysis of species composition of catches in test fisheries. The knowledge of species composition is critical in the estimation of sockeye and pink abundance for the Fraser River Salmon Fisheries Management.

Recent applications of the dual-frequency identification sonar (DIDSON) at Mission and upstream locations have provided a unique opportunity to tackle the problem of species identification. The DIDSON sonar yields high resolution images of individual fish, which provide rich information about the shape, size and behavior of individual fish. Here we report the results of a feasibility study aimed at determining whether DIDSON data collected at Mission can be used to estimate the composition of pink and sockeye. We have collaborated with the Stock Monitoring Group at the Pacific Salmon Commission (PSC) in this project.

We have first performed an analysis of morphometric and behavioral characteristics of individual fish extracted from the DIDSON data collected at Mission from the 2004-2007 management seasons, and then selected a set of feature variables which we believe would be effective in discriminating pink and sockeye. We have investigated several methods for estimating the pink and sockeye composition based on these feature variables, and found that the discrimiant function analysis (DFA) method was most reliable. The performances of the DFA method were evaluated via numerical simulation, and it was found that even though the error rate of this method is not small (about 20%), the averaged bias of the composition estimates is 1% - 3%. We have applied the DFA method to two periods of continuous hourly data (8 hours in period 2007-08-19, and 27 hours in period 2007-08-25/26). The results show that the composition can be highly variable from hour to hour, revealing dynamic passage patterns of different species. The averaged pink/sockeye composition estimates in these two periods are 0.27/0.73 for period 2007-08-19, and 0.79/0.21 for period 2007-08-25/26. This seems consistent with typical salmon migration patterns at Mission, in which the majority of pink salmon arrive in late August and early September. In summary, our work has clearly demonstrated that it is feasible to estimate the pink/sockeye composition at Mission based on DIDSON data. It is also possible that this approach can be applied to similar problems where DIDSON systems are in place. We also recommend research areas for future work.

2 INTRODUCTION

The hydro-acoustics program operated by the Pacific Salmon Commission provides estimates of the daily passage of salmon at a site located near Mission B.C. The focus for Fraser River Panel management is sockeye and pink salmon and the estimates of daily passage for these key species are determined through an analysis of species composition of catches in test fisheries. The knowledge of species composition is critical in the estimation of sockeye and pink abundance for the Fraser River Salmon Fisheries Management.

The proportion of sockeye in estimates of daily salmon passage has historically been derived from the proportion of gilled and girthed sockeye in the total gilled and girthed salmon catches from a gillnet test fishery located downstream of Mission at Whonnock. The accuracy of this approach depends on the key assumption of equal catchability across all species of interest. Variation in catchability due to differences in fish size is addressed through the use of a multi-panel gillnet in an attempt to ensure that the selectivity of the net is as equal as possible across salmon stocks and species of different sizes. However behavioural factors such as swimming speed, schooling and cross-river and vertical distribution are known to influence species specific catchability. In addition, possible species specific removal of salmon from the net by seals has recently been determined as a source of potential bias in the estimates of species composition derived from these catches. Another limitation of this method is that it does not provide temporal and spatial distributions of daily species composition in the river as the daily test-fishing program only operates in a limited time period and in localized areas.

Hydroacoustic technologies using multiple-frequency or broadband sonar systems (Horne, 2000) have shown some potential in the estimation of species identification. While such systems may prove successful in the future, at the present time the underlying technologies have not reached the stage of practical application. Alternatively, recent applications of the dual-frequency identification sonar (DIDSON) at Mission and upstream locations (Xie et. al., 2005; Holmes et. al., 2005) have provided a unique opportunity to tackle the problem of species identification. The DIDSON sonar yields high resolution images of individual fish as they move across the acoustic beams. These images provide rich information about the shape, the size and behavior of individual fish, as well as their behavior as a school. This information, when properly processed, could potentially be used to identify species in the Fraser River at Mission where resident species and migrating salmon display different behaviour and distinguishable distributions of body-length.

Here we report the results of a feasibility study in which DIDSON data collected at Mission was used to estimate species composition. Morphometric and behavioral characteristics of individual fish were extracted from the DIDSON data. We have collaborated with the Stock Monitoring Group at the Pacific Salmon Commission (PSC) in this project, and analyzed DIDSON data they collected at Mission from the 2004-2007 management seasons. Our work has demonstrated that it is possible to estimate species composition based on DIDSON data. In this report, we will describe the data processing techniques and estimation methods, evaluate the performances of the estimation methods we applied, and present preliminary results of applying the methods to the 2007 field data. We also recommend research areas for future work.

3 METHODS

3.1 Technical Background Overview

As indicated above, our approach to species composition estimation is to analyze fish morphology and behavior characteristics extracted from DIDSON sonar image data. One of the essential requirements for this work is a software tool suitable for processing a large amount of image data generated by the DIDSON sonar. Before the project started, Vitech had developed a software tool to track automatically

individual fish in DIDSON data, enabling us to extract behavioral data (e.g. speed, direction, tortuosity) from a large amount of image data. Vitech has since added utilities to the software to enable measurement of individual fish length, which is a key variable to distinguish different fish species. These added measurement tools include user-driven tools and automatic capability, and will be described below.

To perform species identification, we need to construct a set of feature variables (descriptors) from morphometric and behavioral characteristics derived from DIDSON data. Optimal selection of feature variables will require careful analysis and experiment. Once a set of feature variables is selected, we apply classification methods to individual observations and obtain an estimate of species composition. One of the most commonly used classification methods is Discriminant Function Analysis (DFA). It has previously been applied to species identification of fish schools based on echograms (Haralabous and Georgakarakos, 1996; Lu and Lee, 1995), and will be discussed in detail later. We will also discuss the application of an algorithm called Expectation Maximization that does not classify individual observations but allows estimation of species composition directly from a mixture of data from different species.

3.2 DIDSON Sonar

The Dual-frequency Identification Sonar (DIDSON™), developed by SoundMetrics Corp, is a high resolution imaging sonar system, operated at mega-hertz frequencies (1.8MHz or 1.1MHz) (Belcher et al., 2002). The system has a large azimuthal composite beam of 29^0, which consists of 96 (1.8MHz) or 48 (1.1MHz) fan-shaped narrow beams. Each fan beam has an angular resolution of 0.3^0 (or 0.6^0) x 12^0. The spatial resolution depends on range and frequency. For example, at 1.8MHz and a maximum range of 10m, the system has a range resolution of 2cm and an azimuthal resolution of 5cm at 10m. Such resolutions allow the system to generate 2D image with near-video quality.

The DIDSON system is capable of capturing images at a rate up to 21 frames per second. The captured images can be played back like a movie. In most cases, migrating fish (or other moving targets) can be visually identified in the movie. Visual counting of fish and manual measurement of fish size is possible when dealing with a small amount of data. Yet these manual operations become very labor intensive when processing a large amount of data, and in this case, the ability to perform these tasks with some automation is highly desirable.

3.3 Target Tracking Software

In this project, we have used a self-contained software package (IntelliHAT™, **Intelli**gent **H**ydro-**A**coustics **T**racker) which we designed and developed to track individual moving targets recorded by Hi-resolution sonar, such as DIDSON™ (Sound Metrics Corp). The software consists of three components: identification of individual targets from raw images based on cluster analysis, multi-target tracking of identified targets, and subsequent filtering of track data based on pattern recognition. Depending on data quality (judged by the signal-to-noise ratio), raw image files may be preprocessed before target identification. Raw images or preprocessed images are analyzed using clustering algorithms (Theodoridis and Koutroumbas, 2003) to identify potential targets. The positions of the identified targets in individual frames are then fed to a multiple target tracker (Blackman and Popoli, 1999) which will sort the input data to generate track data. The output track data record the spatial position and acoustic intensity of each individual tracked target as a function of time. Unwanted tracks may be removed by applying a pattern recognition filter to the track data.

3.4 Development of Fish Length Measurement Tools

A key technical issue in this project is accurate measurement of fish size which, when combined with other information, can be used in species recognition. Several tools of size measurement have been developed in this project, including

- A user-driven measurement tool that allows the user to connect pixels on an image that are perceived to be associated with a fish.
- A tool that allows the user to draw a rectangular box surrounding a fish on an image and then measure the length of the fish. A simple automatic thresholding algorithm (Gonzalez and Woods, 2002) has been implemented to determine automatically an intensity threshold. Pixels in the box with intensity above the threshold are considered to be due to the fish. The fish length is then calculated from the selected pixels.
- Automatic measurement of target length from a track file. To use this tool, a DIDSON image file is first tracked to generate a track data file which records the time and position of each individual track. Then the user may review the track file and select a set of tracks for size measurement. Alternatively, an algorithm has been implemented to allow automatic selection of tracks that are relatively isolated from its neighbors (The reason for selecting isolated targets is that when fish are too close to one another, it is very difficult to measure their length independently, at a reasonable accuracy. However, in practical applications, selection of relatively isolated tracks may bias species composition estimation, if fish of one species tend to swim closer together than fish of other species.). Then another algorithm examines the image data associated with each selected track and then measures the target size at the frames in which the target is tracked.

3.5 Optimal Estimation of Live Fish Length

Automatic measurement of live fish length in the field is much more challenging for a number of reasons. First, fish in tight aggregations are difficult to measure unambiguously. Second, fish often flex as they move across the acoustic view, making it difficult to determine which measurement in time best represents the length. Fish may also be at some locations and orientations that make measurements difficult or less accurate. Despite these limitations, chances are reasonably good that we could derive a reasonable estimate of fish length from a sequence of measurements over frames, as long as local fish density is not too high. Moreover, here we are more interested in collecting statistical distributions of fish length for species composition estimation, than measuring each individual fish with high precision.

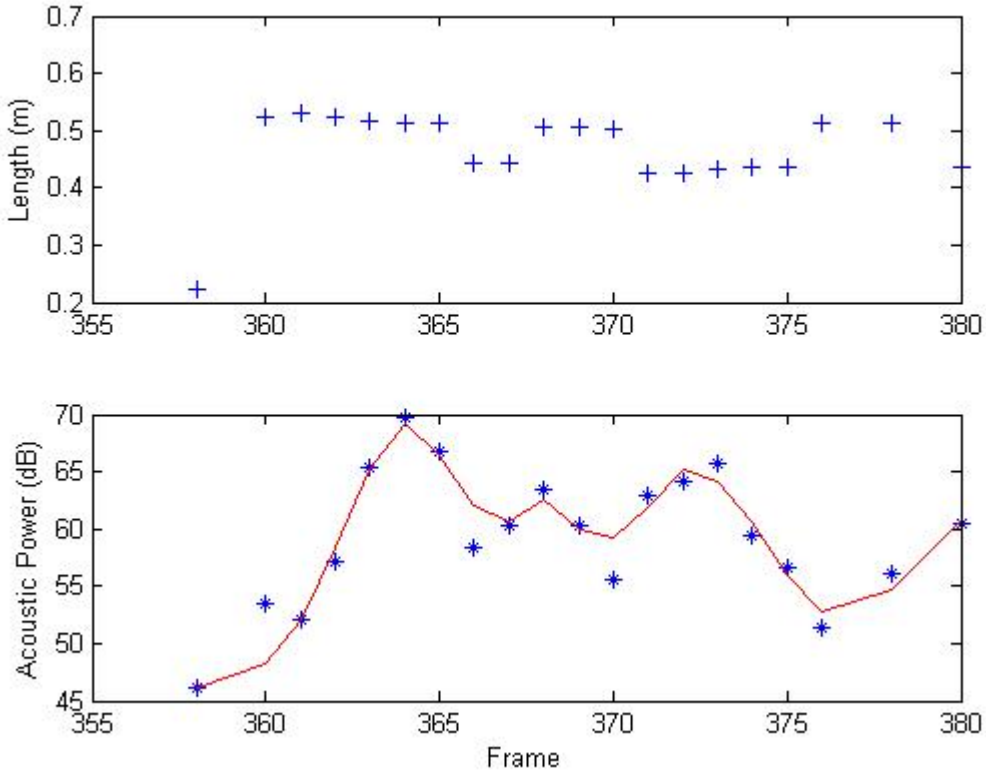

Fig. 1.: Length measurements of a live fish as a function of time (frames). The lower plot is the corresponding 'brightness' of the fish image (see the text for more explanation).

Figure 2. **2.**(upper plot) shows an example of size measurements of a fish as it swims across the acoustic beam, where the horizontal axis is frame number. These measurements are time dependent, and can be highly variable. Now the question is which measurement represents the best estimate of the fish length, given the variability. Intuitively, we can say that a measurement is the best estimate when the fish image is brightest. Therefore, we define a 'brightness' parameter, which is the total acoustic intensity in an 'imaged' fish, normalized by the number of the pixels associated with the fish. The brightness corresponding to the length measurements is shown in the lower plot of Fig. 2.. As can be seen, the brightness can also be highly variable, but we smooth the data using a robust curve-fit algorithm, as shown by the red line in the plot. Then we take the length measurement corresponding to the peak brightness (after smoothing) as the length estimate for the fish. This simple approach may not be optimal, but it does generate reasonable length distributions from field data, as seen below.

3.6 Derivation of Other Characteristics of Fish from DIDSON Data

In addition to fish length, we can also derive other characteristics of individual fish from DIDSON data. As mentioned before, the output track data from the software record the position and acoustic intensity of each identified target over frames, and allow us to obtain fish behavioral characteristics. One of these is path-averaged speed, which is the length of a target trajectory divided by the time it spans. This represents how fast a fish can swim on average, regardless of the direction. Another characteristic is cross-beam velocity (speed and direction), which provides an indicator of the upstream/downstream direction for migrating salmon.

Another characteristic related to an imaged fish is the aspect ratio of the image. This is defined as the ratio of standard deviations of the pixel positions in two perpendicular directions. Typically, we choose one direction to be the head-to-tail direction (we use principal component analysis to determine the two directions automatically). In our definition, the smaller the aspect ratio is, the more elongated the imaged fish is. In general, if a fish larger and brighter, it is more likely to have a smaller aspect ratio in its image.

As mentioned above, fish length is measured when the 'brightness' of an imaged fish is maximum during its lifetime. Brightness is defined as the total acoustic intensity in an 'imaged' fish, normalized by the number of the pixels associated with the fish. Note that the acoustic intensity is not calibrated, but compensated for spherical spreading of sound. Brightness is related to the acoustic scattering property of a fish. Since this parameter is not calibrated and thus is system-dependent, it should be used with care.

3.7 Feature Selection for Species Classification

The goal of feature selection is to identify a set of feature variables that are the most effective in discriminating objects in different classes of interest. The first step of feature selection is to identify what feature variables should be derived from measurement data, and this can be guided by our understanding of what would really separate the classes under investigation. We can then test separately each of these features, by using a measure of discrimination power. The second step is to consider combinations of features and then use the same measure of discrimination power to select the best combination. However, the number of possible combinations increases rapidly as the number of available feature variables increase. Feature selection itself is an area of extensive research in pattern recognition, and in this project, we did not intend to conduct an exhaustive search of optimal feature variables. Instead, we derived a set of feature variables that we consider may be relevant to our problem, as those mentioned above. We then included them as long as the overall discrimination power kept increasing.

One measure of discrimination power is the so-called divergence, which is related to the Bayes classification error. If samples are assumed to be normally distributed, the divergence between class *i* and class *j* is given by

$$d_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'(\mathbf{S}_i^{-1} - \mathbf{S}_j^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + trace\left\{\mathbf{S}_i^{-1}\mathbf{S}_j + \mathbf{S}_j^{-1}\mathbf{S}_i - 2\mathbf{I}\right\},$$

where $\boldsymbol{\mu}$ and $\mathbf{S}$ represent the mean vector and covariance matrix in the normal distribution (Theodoridis and Koutroumbas, 2003).

It is also possible to transform a set of feature variables into another set of variables which is considered optimal under a certain criterion. For example, consider a definition of between-class scatter matrix

$$\mathbf{R}_b = \sum_{i=1}^{G} P(\omega_i)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)',$$

where $P(\omega_i)$ is prior probability class *i*, and *G* is the number of classes. $\boldsymbol{\mu}_i$ is the mean vector of feature variables in class *i*, and $\boldsymbol{\mu}_0$ is the mean vectors of all feature variables defined as

$$\boldsymbol{\mu}_0 = \sum_{i=1}^{G} P(\omega_i)\boldsymbol{\mu}_i \ .$$

For classification, it is desirable for this matrix to be 'large' compared to the covariance of feature variables within individual classes. Thus we may introduce another matrix

$$\mathbf{D} = \mathbf{R}_c^{-1}\mathbf{R}_b,$$

where $\mathbf{R}_c$ is the mean of the covariance matrix within individual classes:

$$\mathbf{R}_c = \sum_{i=1}^{G} P(\omega_i) E\left[(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)' \right].$$

We can find a linear transformation of a set of feature variables $\mathbf{x}$, $\mathbf{y} = \mathbf{C}^{\mathrm{T}}\mathbf{x}$, such that the sum of the diagonal elements of matrix $\mathbf{D}$ for $\mathbf{y}$ is maximized. It turns out that the transformation matrix $\mathbf{C}$ contains the eigenvectors corresponding to the K nonzero eigenvalues of matrix $\mathbf{D}$ for $\mathbf{x}$ (Theodoridis and Koutroumbas, 2003).

3.8 Estimation Methods for Species Composition

The ultimate goal of this project is to estimate species composition based on DIDSON sonar observations. Estimation methods can be categorized into two groups. One is based on species classification, in which individual observations are analyzed to determine which species they may have originated from. Species composition is then estimated by counting the number of observations classified to each species and performing appropriate corrections. The other group of methods does not classify individual observations. The information on which species an individual observation is associated with is unknown without performing classification, but the uncertainty can be averaged out, allowing extraction of the composition and other model parameters (such as mean and standard deviation) from a set of observations.

We employed two classification-based methods in this project. One is Discriminant Function Analysis (DFA), which is based on the assumption of normality and Bayesian Decision theory. In this approach, the probability distribution of samples is assumed to be a multivariate normal distribution, and the parameters of the normal distribution are estimated from training samples. Classification is then based on Bayes Decision Theory, which classifies an observation into a group, if the posterior (or a posteriori) probability for this group is the maximum (see the appendix for more details).

After DFA is trained and applied to a classification problem, the classification result needs to be corrected for classification errors. Let p_{ij} be the probability of an object in Group i being classified to Group j, and m_j be proportion of objects classified to Group j. Then we find (McLachlan, 1992)

$$E\{m_j\} = \sum_{i=1}^{G} \pi_i p_{ij},$$

where π_i is the true proportion of objects in Group i. Note that p_{ij} is a conditional probability on a specific set of data. This can be obtained as we train the DFA (or ANN) classifier. If we estimate π_i based on one realization, then the estimate can be obtained by solving a set of linear equations:

$$\hat{\boldsymbol{\pi}} = \mathbf{J}^{-1}\mathbf{m}$$

$$\mathbf{J} = \begin{bmatrix} p_{11} \cdots p_{G1} \\ \cdots \\ p_{1G} \cdots p_{GG} \end{bmatrix}$$

where $\mathbf{J}$ is the classification matrix. However, for a single realization, the estimate can be out of bounds (0-1). A simple solution is to set it to the closer bound.

Here we also introduce an estimation method that does not classify individual observations. This method uses an algorithm called Expectation Maximization (EM). For simplicity, we consider a two mode mixture model, where the total probability density function (PDF) of observations is given by:

$$f(\mathbf{x}) = P_1 f_1(\mathbf{x}) + P_2 f_2(\mathbf{x}),$$

where $f_1(x)$ and $f_2(x)$ are the PDF for each mode, and P_1 and P_2 are the composition (typically, these probability distributions are assumed to be normal). The information on which species a feature sample is associated with is unknown without performing classification, but the EM algorithm handles this problem by introducing an unknown parameter, j_k to represent the species index from which the k-th sample is drawn. EM then defines a likelihood function and averages the function over j_k (Theodoridis and Koutroumbas, 2003). The composition and other model parameters (such as mean and standard deviation) are derived by maximizing the likelihood function (see the appendix for more details).

3.9 Performance Evaluation of Estimation Methods

After an estimation method is selected, an important step is to evaluate its performances in terms of classification errors and overall estimation bias and variance. A straightforward evaluation approach is to collect a set of monospecific samples, independent of those used for training purposes, and test the output of an estimation method against the actual input. However, the set of monospecific samples at our disposal is often finite and has to be used in both training and testing. One evaluation approach in the case of a limited number of training data is the bootstrap method, which is commonly used in pattern classification (Han and Kamber, 2005). This method generates a set of training data, by randomly and uniformly sampling an available data set with replacement (i.e. the same data may be selected more than once). For a data set of size N, if we select N data points with replacement, there will be roughly 63% of the original data selected, while the rest (37%) will be left out. Those selected will be used for training and those left out used for testing. This procedure is repeated for a number of times, leading to a statistical distribution of the outputs, from which the performances can be evaluated in a statistical sense. Although there are other evaluation approaches recommended for pattern classification such as cross-validation, we feel the bootstrap method is appropriate within the scope of this project.

Using the bootstrap method, we can also evaluate the accuracy of composition estimates generated from a classifier. Typically, we randomly select half of available monospecific data for training and the other half for testing. We then construct from the testing data new datasets of different species compositions. The new datasets are then fed to the classifier and the classification results are then compared with the true compositions.

4 PROCESSING OF DIDSON SONAR IMAGE DATA

4.1 DIDSON Sonar Image Data

The Stock Monitoring Group of the Pacific Salmon Commission (PSC) has provided DIDSON data collected at Mission from the 2004 to 2007 management seasons. These data provide different representative scenarios of species composition and schooling behavior. For example, a preliminary analysis of the 2005 DIDSON data by PSC shows a two-mode distribution of fish length in near-bottom areas and a clear schooling of upstream migrating fish (Xie *et. al.*, 2005). In the 2007 season, pink salmon passed Mission at a significantly high number, providing a good opportunity for testing species identification between pink and sockeye.

In the following sections, we will provide some typical results from processing the DIDSON data at various stages. This should help us better explain the data processing procedure.

4.2 Data Processing by Vitech's Software

Here we show an example of target tracking results from Vitech's software. Figure 2. shows a frame of DIDSON image data collected in 2007, where pink passed through the observation area in a large quantity. The red circles represent the target locations automatically identified by the software. Note that not all visually identifiable targets are identified by the software in every frame they appear; some of them may be identified later (or earlier). As long as a target is identified for a minimum number frames, it will be tracked in the subsequent tracking process.

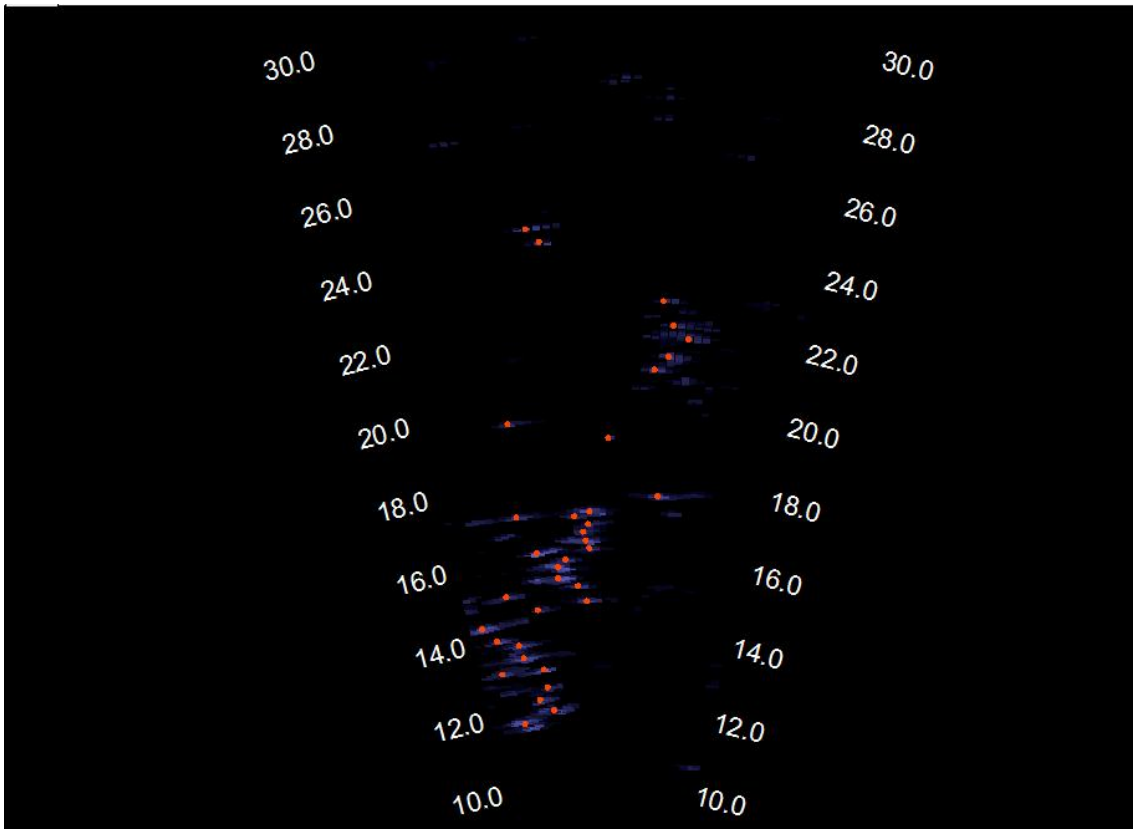

Fig. 2.: A frame of DIDSON image data in the case of a high pink passage collected on September 9, 2007. The red dots represent the locations of identified targets.

The positions of those red dots for each frame are recorded and fed to the multiple target tracker (MTT) based on radar tracking algorithms. The result of MTT is the so called track data, which provides time, position and other auxiliary information of individual targets in all the image frames where they are tracked. Figure 3. is a plot of the track data obtained from the video data in Fig. 2., which shows the resulting tracks in range versus time coordinates. Sixteen different colors are used in a rotational fashion to differentiate individual tracks. Note that this is different from an echogram which only records target passage at one beam angle; the track plot as in Fig. 3. shows target passage through the entire observation area.

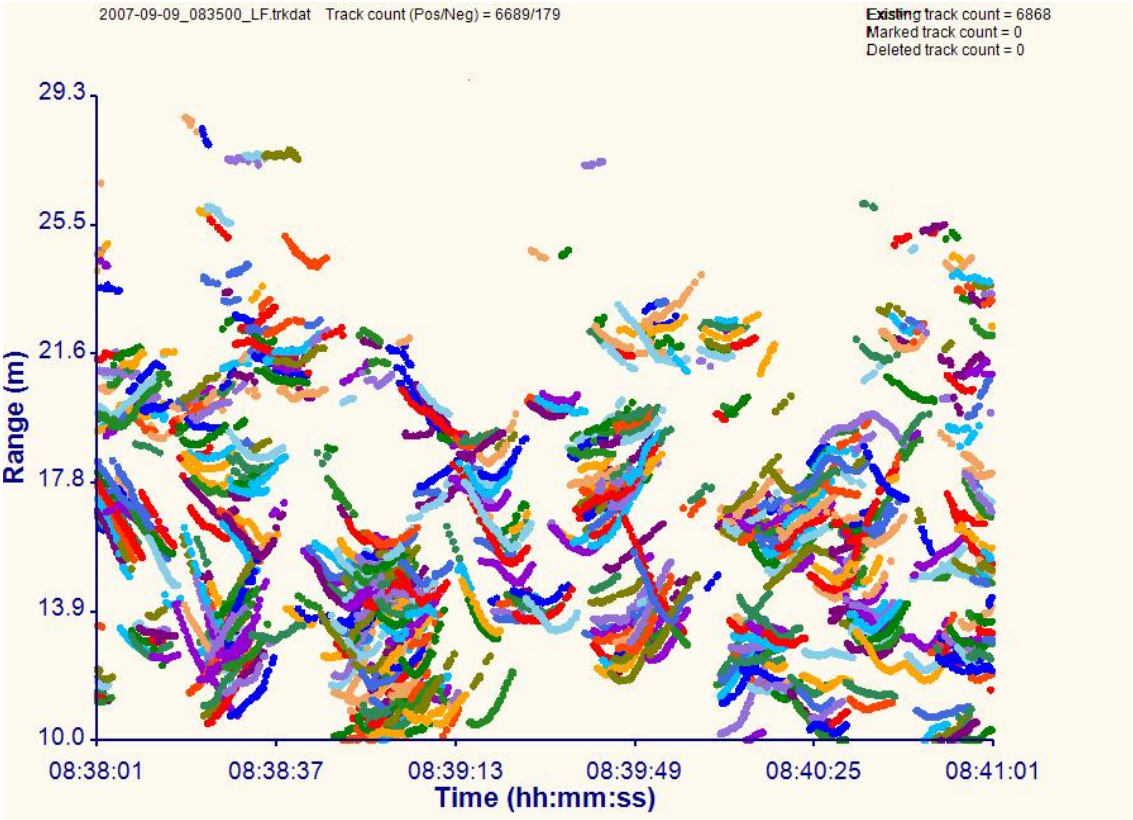

Fig. 3.: Track data generated by the software from the raw image data in Fig. 1. Colors represent different tracks.

4.3 Measurement of Fish Size

As indicated by Section 3, we have developed three fish size measurement tools. The simplest one is a user-interactive manual measurement tool that allows the user to connect, using a computer mouse, pixels on an image that are perceived to be associated with a fish. The PSC group collected some DIDSON image data of known targets, which are used here to test our measurement tools. Figure 4. shows a frame of the data due to a concrete fish of length 61.5cm (tip to tail) placed at range 8.15m. It can be seen that most of the fish body is very bright, but the tail part seems a bit difficult to identify. This suggests that fork length (or post-orbital fork length) is a more measurable parameter than tip-to-tail length.

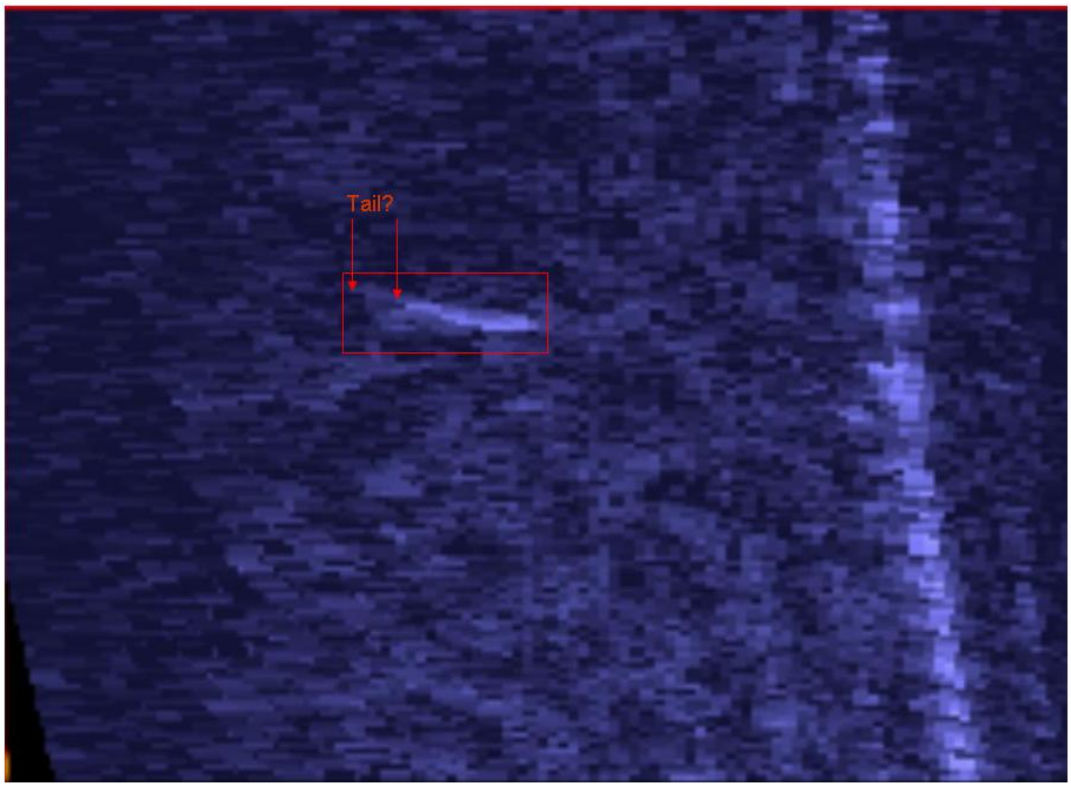

Fig. 4.: Image of a concrete fish (File 2006-06-29_#008_HF.ddf). The red box encloses the fish. It can be seen that the tail is not as clear as the main body.

Figure 5. shows a pixel connection measurement of the fish using IntelliHAT, giving a measured length of 60cm. We measured the target in 20 different frames, and the mean length is 60.6cm with a standard deviation of 1.7cm. Our manual measurements (pixel connection) of these targets are very close to those obtained from the software provided by the DIDSON manufacturer. Despite the variability in the measurements, the mean result seems consistent with the target size.

Our second size measurement tool allows the user to draw a rectangular box surrounding a fish on an image and then measure the fish length based on the pixels in the box. Figure 6. shows an in-box measurement with automatic threshold detection. Comparing with Fig. 5., it can be seen that the tail and tip of the fish were not detected, leading to a measured length of 43.2cm. Therefore, the in-box measurement would more likely correspond to the POF (Post Orbital Fork) length.

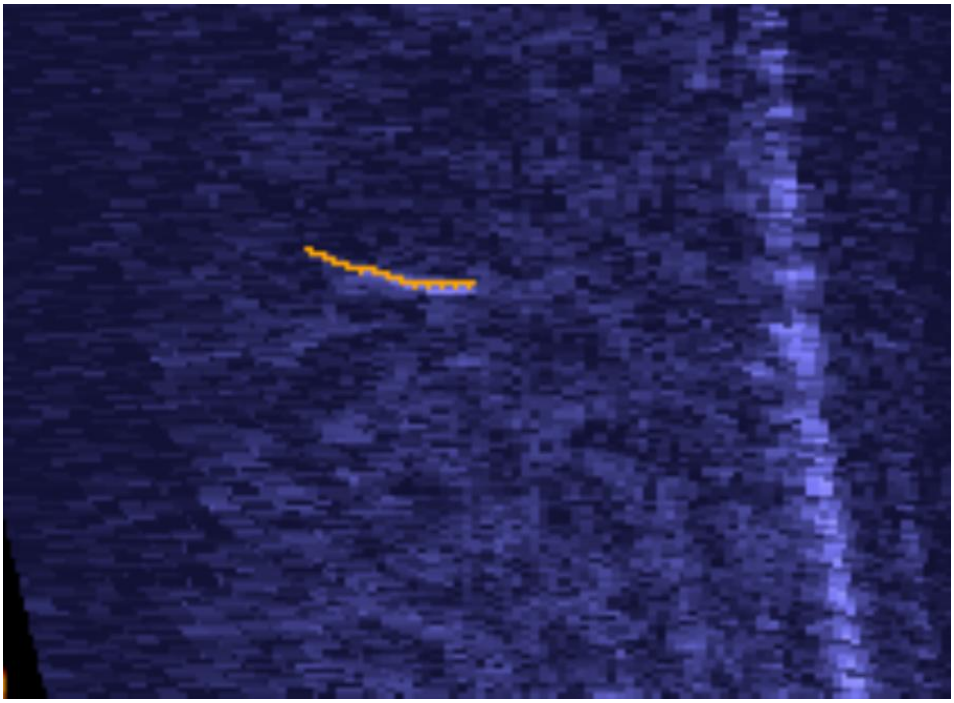

Fig. 5.: Pixel connection measurement of the concrete fish in Fig. 4.. The line connects pixels from tail to tip, yielding a tail-to-tip length of 60cm, somewhat less than the target's tail-to-tip length (61.5cm).

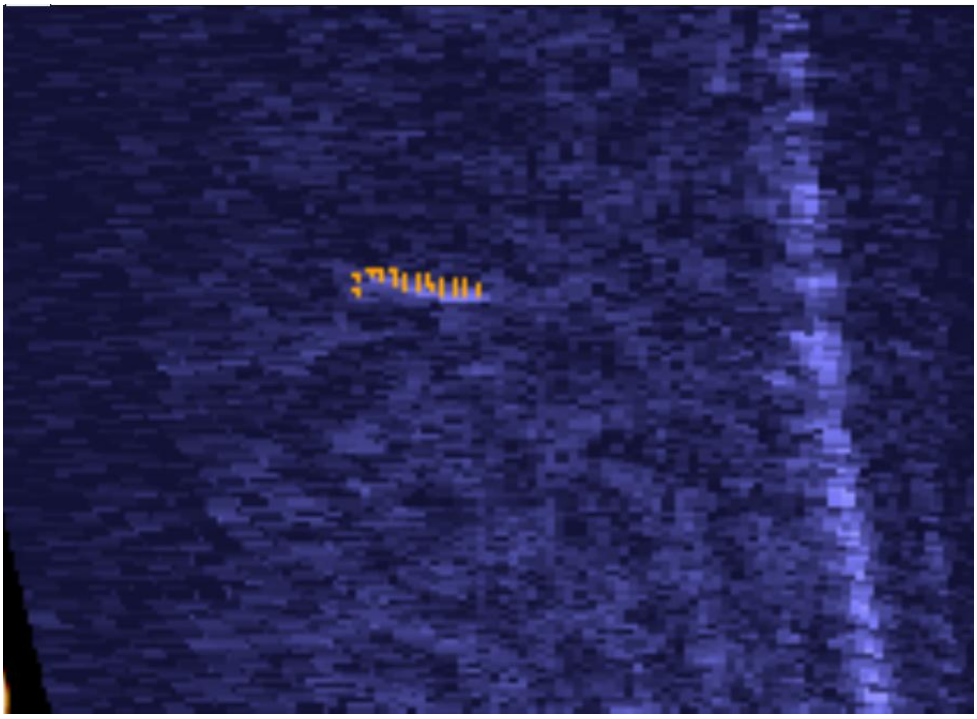

Fig. 6.: In-box connection measurement of the concrete fish in Fig. 4.. The dots represent pixels with intensity above a threshold automatically determined by the software. The resulting measured length is 43.2cm, consistent with the POF length of the target.

Our third tool is an automatic one based on the box measurement tool. To test this tool, we let the software track the same image file as in Fig. 4., and then selected the track of the fish (which in this case is a stationary target). The software examined the image data associated with the selected track and then calculated the target length for 200 frames. Figure 7. shows the histogram of the measured length, where it can be seen that the dominant peak is at 44cm. Although the POF length of the target was not measured, it should be a few centimetres less than the tip-to-tail length. Therefore, the automatic measurement tool would probably yield a result a few centimetres less than the POF length.

The same concrete fish was also measured at range 9.5m, but the bottom interference was quite strong. This makes automatic measurements more uncertain, as can be seen in the histogram in Fig. 8.. This result helps explain why measurement results from field data have a large variability in some situations, as will be seen below.

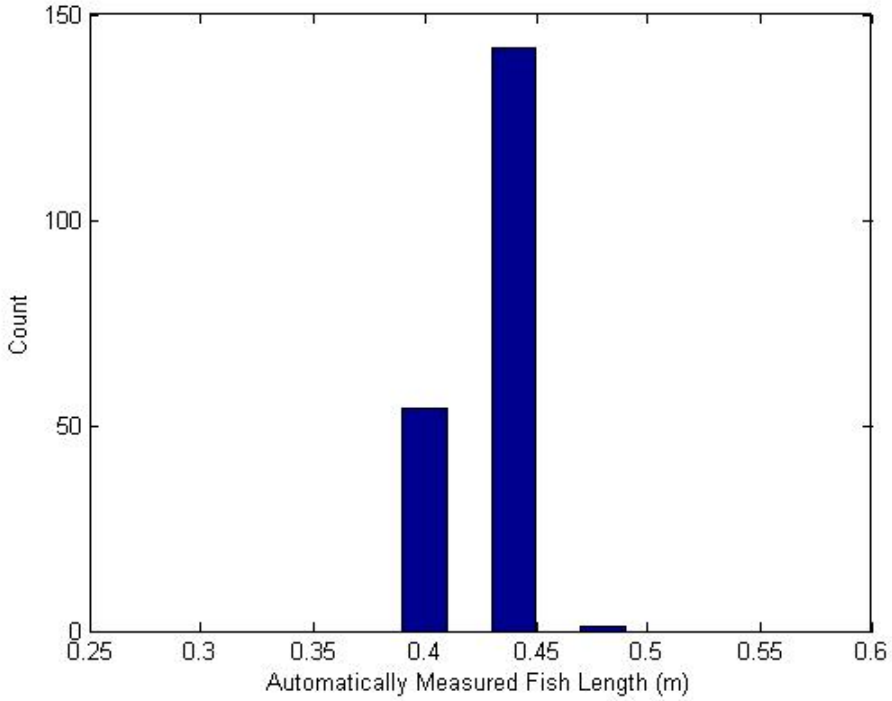

Fig. 7.: Histogram of automatically measured length of the fish in Fig. 4. at range 8.15m, for 200 frames. The dominant peak is within 43-45cm, consistent with the POF length of the target.

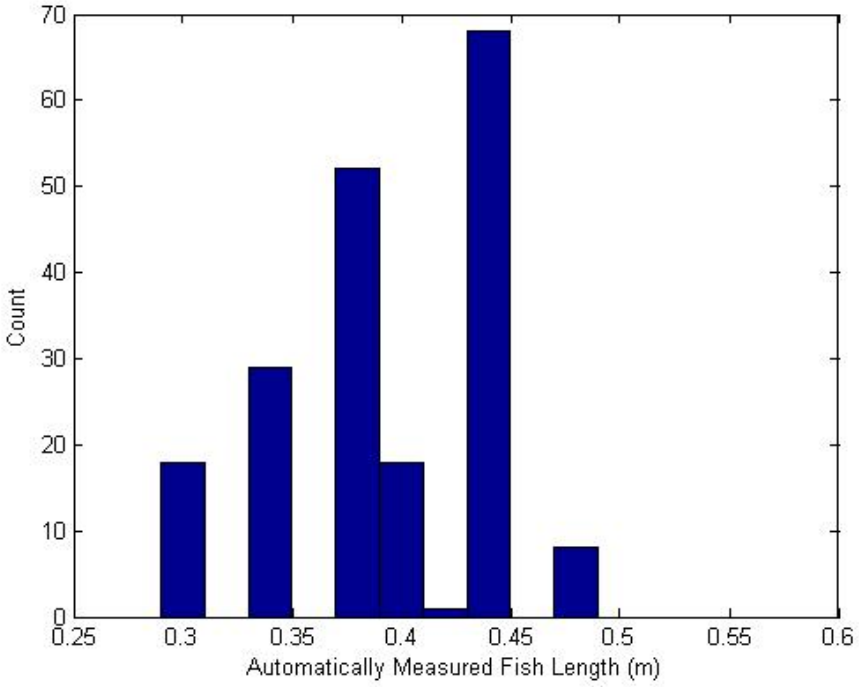

Fig. 8.: Histogram of automatically measured length of the fish in Fig. 4., but located at range 9.5m, for 200 frames. The dominant peak is still within 43-45cm, but the distribution is much broader, due to bottom interference.

4.4 Distributions of Live Fish Length

Now we present some results from live fish data collected at Mission, where a single species (e.g. small fish, pink, sockeye, and chum) dominates, or there is a mixture of different species. For this illustration, we let the software automatically select relatively isolated tracks and calculate the target sizes associated with the selected tracks, except for the chum files, for which we used manual measurement. We define a track as isolated from its neighbours if the distance from the track to its nearest neighbour is greater than 1-2 times their cluster size. This ratio allows us to balance the track selection between not eliminating too many tracks and avoiding closely adjacent tracks.

Figure 9. is a case where small fish dominate. Visual inspection of the data shows very few larger fish (sockeye) passing through. As expected, the length distribution peaks around 20cm, which is substantially less than the pink or sockeye peak as in the following figure.

Figure 10. shows a series of distributions from newly collected data (2007-08-07, 20007-08-19, 2007-08-25, 2007-09-09). In Fig. 10.(a), we can see 3 visible peaks at 20cm, 35cm, and 45cm, and the 45cm peak appears quite strong. In the subsequent plots, the 45cm peak gradually diminishes. These plots correspond to a period in which pink and sockeye co-migrated and later sockeye gradually disappeared and pink became dominant. We may assume that the 35cm peak is associated with pink, and the 45cm peak with sockeye. The dominant length of pink and sockeye measured by the software is significantly less than the test fishing data collected at Johnstone Strait (48.8cm and 54.4cm), but the difference seems consistent. However, the distributions of pink and sockeye have a long tail on the right side, with measured length extended well over 1m. This is due mostly to the effects of adjacent fish and background interference, and is an area of improvement.

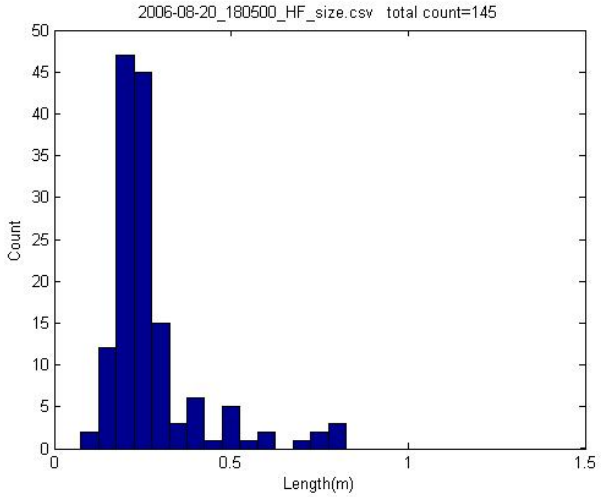

Fig. 9.: Histogram of length measurements of a DIDSON image file (2006-08-20_180500_HF.ddf), where small fish dominate. Note the peak at 20cm.

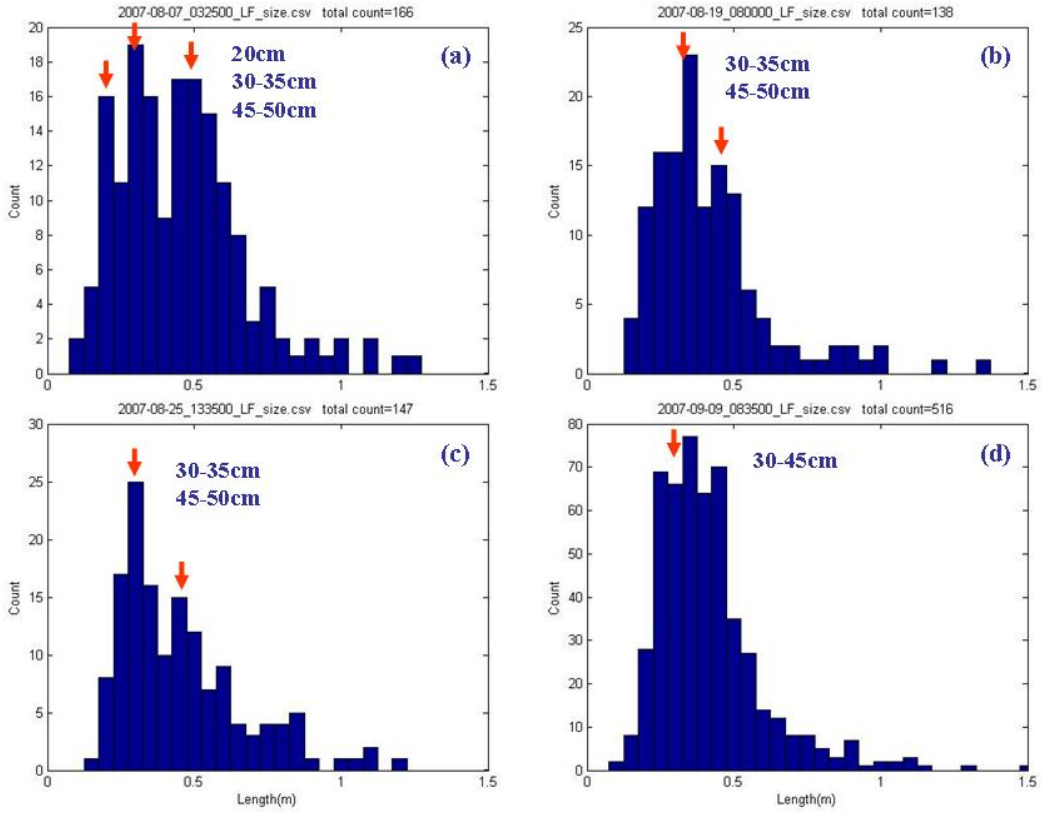

Fig. 10.: Histograms of length measurements from 2007 Mission DIDSON data collected on four days. Each data set has a time span of 20min. (a) 2007-08-07; (b) 2007-08-19; (c) 2007-08-25; (d) 2007-9-09. Note that the peak at 45-50cm in plot (a), (b) and (c) gradually diminishes, and that it disappears in plot (d). These plots correspond to a period in which pink and sockeye co-migrate and later sockeye gradually disappear and pink become dominant.

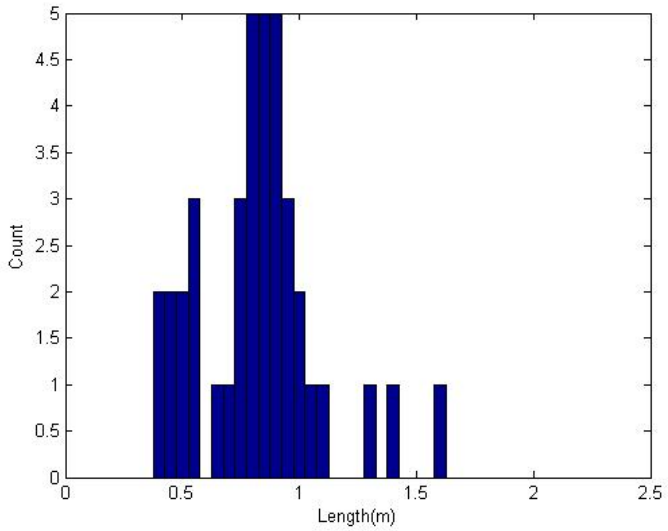

Fig. 11. : Histogram of manual measurements of fish length from two DIDSON image file (2006-09-18_133000_LF.ddf and 2006-09-19_043000_LF.ddf) where chum dominate (39 fish in total). Note the peak at 80-90cm.

The chum image files are a challenge, because cross-beam interference (mainly due to bistatic scattering of a target, which lightens up fan-beams not directly ensonifying the target) generally biases measurements upwards. Therefore, we selected a few tracks (39 in total) and manually measured their length, as shown in Fig. 11.. The dominant group has a peak at 80-90cm. There is also a secondary group which is concentrated around 50cm. This may be due to pink or sockeye. There are a few measurements well above 1m, which are due probably to the cross-beam interference making it difficult to judge the boundary of fish images, even visually.

4.5 Selection of Training and Testing Data

One of the important tasks in this project is to select sufficient monospecific samples to train and test classifiers. We used test fishing information to find time periods in which a single species dominates. Test fishing data were usually collected 20-km downstream of the sonar observation site. Within the distance, we do not expect species composition to change significantly. We then collected samples from data in the time periods in which single species were expected to dominate. Even during these periods, there were occasionally local small fish or other salmon species in the data, and we carefully reviewed the image data to exclude fish that did not appear to belong to the species of interest. For example, local small fish look smaller and directionless in video playback of the data, and can be easily identified and excluded from pink or sockeye samples. Pink tend to swim tightly in groups of two or more, and when selecting pink samples, we focused on those targets in groups.

Ideally, these samples should have been taken under similar environments with similar hardware configurations. We have examined data from the 2005-2007 seasons and decided to use the 2007 data, because a significant part of the 2007 data were collected under similar conditions, providing sufficient samples for training, testing and application.

5 RESULTS and DISCUSSION

5.1 Feature Variable Analysis

In our application, there are three categories of characteristics that could be used to construct feature variables: 1) individual fish behavior such as swim speed and direction; 2) morphological data of individual fish such as length and shape; 3) morphological and behavioral characteristics of fish aggregations or schools. Although all these characteristics may be extracted from image data, we focus on those that are considered most relevant to our problem.

Intuitively, the most obvious feature available from the image data that distinguishes the species of salmon is fish length. Since pink and sockeye are only different by about 10cm in length on average, their distributions of length are expected to overlap. Figure 12.(a) shows the histogram of measured fish length in the training data set, where it can be seen that the peak length is around 30cm for pink (upper pane) and 50cm for sockeye (lower pane). Since the distributions are fairly broad, there is considerable overlap between these two distributions (the long tails in these distributions were mostly due to measurement ambiguity when fish are too close to each other).

Another feature that seems to be able to distinguish the two species is acoustic backscatter intensity (brightness as defined in Section 3), whose distributions are shown in Fig. 12.(c). The pink data have a much broader distribution with the peak in 60-65dB, while the sockeye data have a more narrow distribution peaked in 65-70 dB. The difference in the peaks in the distributions suggests that acoustic reflection coefficient of fish body is different for pink and sockeye. Note that since the sonar system is not calibrated, the value of acoustic intensity depends on the system configuration and should be used only in data with the same configurations as the training data.

The path-averaged swim speed distributions, defined in Section 3, are plotted in Fig. 12.(b). The pink data have a narrower distribution with a peak between 80-90 cm/s, while the sockeye concentrate around 70-80 cm/s with a broader distribution. Another feature of some interest is the aspect ratio of a fish body as measured on an image (see Section 3 for definition), and Fig. 12.(d) shows its distributions. The smaller this value, the more elongated a fish body is. Of course, at low signal-to-noise ratio, it would be difficult to measure this parameter accurately. As seen in Fig. 12.(c), the sockeye data have stronger acoustic intensity, and as a result, the aspect ratio distribution for sockeye is much narrower around 0.2, while the pink data have greater uncertainty (broader distribution).

The discrimination measures mentioned in Section 3, such as divergence, can be used to test the discriminating power of these features. If we test each individual feature, acoustic intensity and aspect ratio have the largest divergence (best discriminating power), followed by length and swim speed. If we test all the features together, the divergence is slightly larger than the sum of the divergence of each individual feature, suggesting that none of these features have a negative effect on the overall divergence.

From these features, we can also construct a canonical variable (see Section 3), and then calculate the divergence of the canonical variable. After some investigation, we found that the divergence is the best when length, acoustic intensity, and swim speed are used to define the canonical variable. Use of all the features does not really improve the divergence. Figure 13. shows the distribution of the resulting canonical variable for both pink (upper pane) and sockeye (lower).

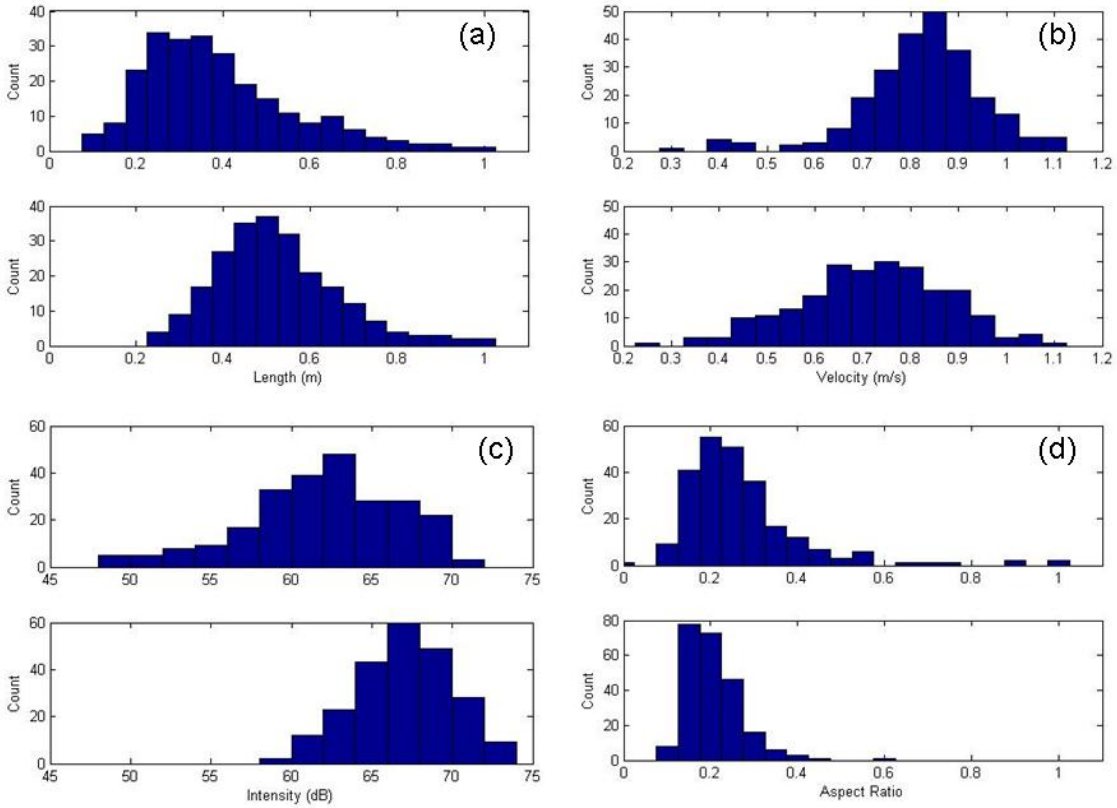

Fig. 12.: Histogram of feature variables of the training data set. (a) fish length; (b) Swim speed; (c) Acoustic intensity; (d) Aspect ratio of fish body. The upper pane is for pink and the lower one for sockeye.

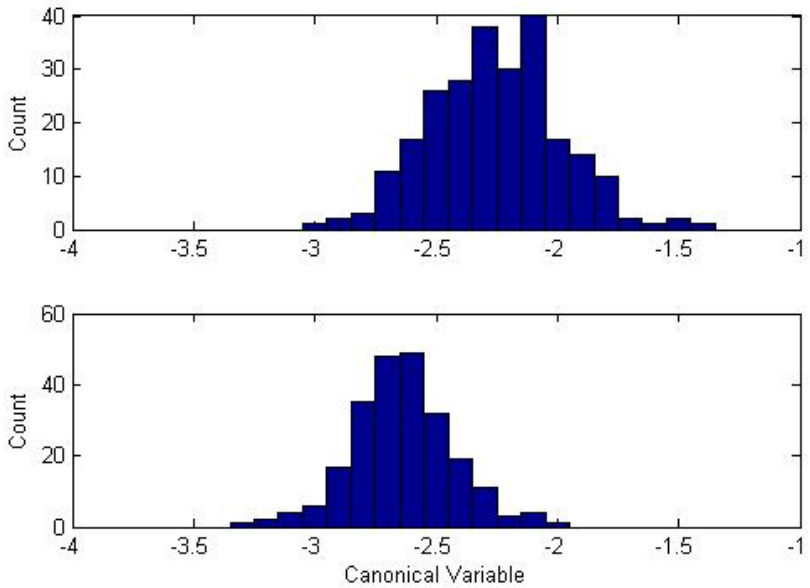

Fig. 13.: Histogram of canonical variable defined by fish length, swim speed, and acoustic intensity. The upper pane is for pink and the lower one for sockeye.

5.2 Training and Evaluation of the DFA Classifier

As indicated in Section 4, there are potentially four species in the DIDSON data, namely, small fish, pink, sockeye and chum (noise would be an additional category). Although we have presented fish length distributions from small fish data and chum data in other years, we did not collect sufficient samples of small fish or chum from the 2007 data to training the DFA classifier. So in this project, we trained the DFA classifier using only pink and sockeye training samples collected in 2007. Composition estimation could be biased without including other species and noise as additional categories, and this will be discussed later.

We use the four feature variables in Fig. 12. and cross-beam swim speed, to train the DFA classifier. Cross-beam swim speed is also an indicator of whether a target is moving upstream or downstream, and could be used to discriminate against small fish or noise. We include this parameter mainly for the purpose of future work, which may be expanded to include additional categories. Having selected these feature variables, we use the bootstrap method described earlier to evaluate the performance a DFA classifier. Figure 14. shows the distributions of the rate of misclassification resulting from 1000 realizations, where the upper plot is for the rate misclassification of pink as sockeye and the lower plot is just the opposite. As described in the appendix, the prior probabilities in the DFA classifier are chosen to balance the rate of misclassification, and therefore we can see that the misclassification rate is similar in both cases, with a mean of 0.19 and a standard deviation of 0.04.

We further evaluate the ability of the classifier to estimate species compositions of pink and sockeye, using the simulation method described in Section 3. Figure 15. is the distribution of the estimated pink proportion resulting from 100 realizations, where the true pink proportion is about 27% of the total population (the sockeye proportion is 63%). The mean estimate is 28% with a standard deviation of 7%. See Table 1 for the DFA estimates of different pink proportions in the pink/sockeye composition. It is seen that even though the error rate of the classifier is not small, the bias of the composition estimates is quite small (1% - 3%), because the error rate is balanced between the two species. The standard deviations of these estimates range from 6% to 8%. It is reasonable that the standard deviation is higher when the proportion of one species is very small, than when the proportions are close.

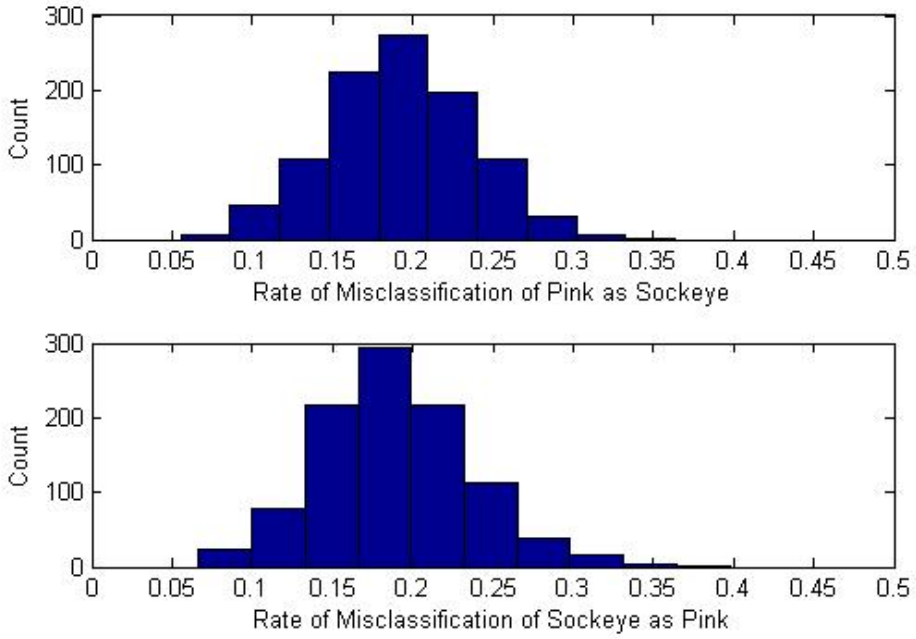

Fig. 14.: Histogram of the rate of misclassification of a DFA classifier. The upper pane is the rate of misclassifying pink as sockeye, and the lower pane is the opposite.

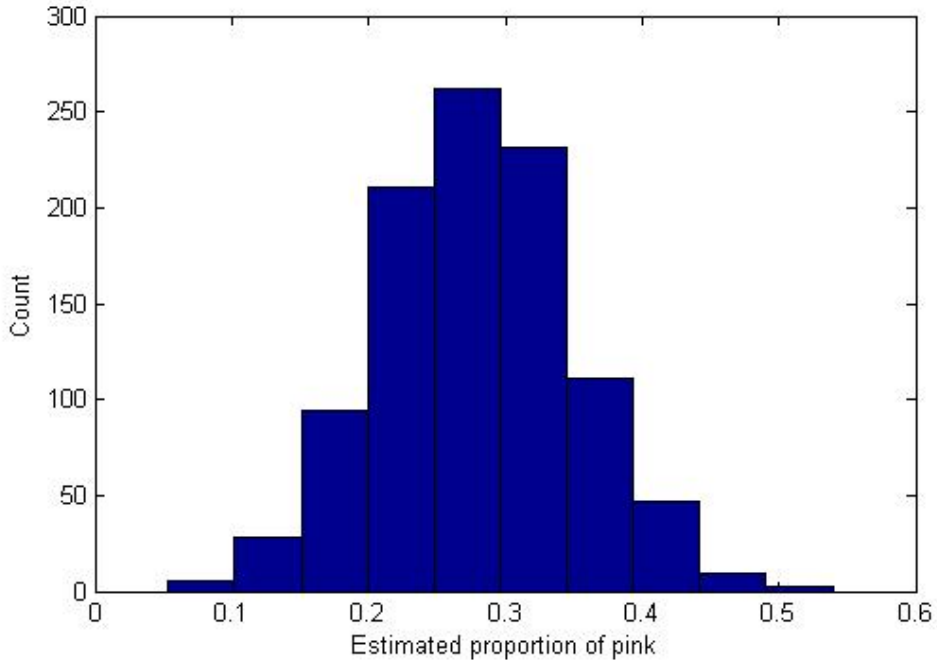

Fig. 15.: Histogram of pink proportion estimated by the DFA classifier. The mean is 0.28 and the standard deviation is 0.07. The true value is 0.27.

True Pink Prop	Mean DFA Est.	STD DFA Est.	Mean EM Est.	STD EM Est.
0.05	0.08	0.084	0.43	0.055
0.14	0.16	0.076	0.35	0.069
0.27	0.28	0.07	0.41	0.063
0.38	0.39	0.064	0.45	0.055
0.49	0.49	0.062	0.5	0.048
0.63	0.63	0.066	0.57	0.041
0.75	0.74	0.071	0.62	0.046
0.87	0.85	0.076	0.67	0.044
0.95	0.93	0.079	0.63	0.059

Table 1: DFA and EM estimates of different pink proportions.

5.3 Evaluation of the EM Algorithm

The ability of the EM algorithm to estimate species composition can be evaluated by following the same approach as described above. The results are given in Table 1. For the EM algorithm, we found that use of the canonical variable as shown in Fig. 13. yields better results than the use of the four feature variables in Fig. 12.. It can be seen that although the EM algorithm generates smaller variance in the estimates comparing than the DFA classifier, the bias is substantially larger when the true proportion is relatively small or large. In fact, when the true proportion is 0.05 or 0.95, the EM estimates break down. This is related to the nonlinearity in the algorithm. Since the EM algorithm does not yield as good performances as the DFA classifier, we will not apply it to field data.

5.4 Application of the DFA Classifier to Field Data

Having evaluated the DFA classifier, we can now use it to estimate species composition in some of the 2007 Mission data. In this preliminary study, we chose data collected under system configurations similar to those for the training data. In these configurations, the DIDSON sonar is operated at the lower resolution mode, covering ranges from 10m to 30m, with a frame rate of 4-5 frames per second. Two time periods were selected: one is the day of 2007-08-19 where sockeye were dominant, and the other is the days from 2007-08-25 to 2007-08-26, where pink were becoming more significant. We happen to have continuous hourly data (each dataset has a time length of 20-25 minutes) in these two periods.

As described in Section 4, we first tracked all targets in these datasets and generated corresponding track files. We then examined the track files, and eliminated obvious noise tracks and tracks due to milling fish. After the cleaning, we allowed the software to measure the fish length associated with every track. We did not purposely select isolated tracks for this analysis. Although this makes it more difficult to measure fish length in a tight group of fish, we feel that selecting relatively isolated tracks would underestimate the proportion of pink, since pink salmon tend to swim in groups.

Figures 16. and 17. show the results of species composition estimates over time for the two time periods, together with fish passage rate, where we have normalized the total count in each dataset to hourly passage rate for comparison. These plots show that the composition can be highly variable from hour to hour, revealing dynamic patterns of fish passage of different species. For example, visual inspection of the raw data suggests that the peak passage rate at Hour 7 in Fig.16. is caused by the

arrival of a number of large groups of pink, boosting the pink proportion from 17% to 33%. Pink have arrived in a significant number one hour before the peak (Hour 6), but are probably outnumbered by sockeye. This is why the pink composition drops despite the increase of the total passage. One hour after the peak (Hour 8), the total passage drops, but the pink proportion continues to increase, indicating that more sockeye than pink have already passed the observation area. Another feature to notice in Fig. 16. is that the total passage is the lowest at Hour 2 with the pink proportion up to 70%. However, we could not find pink in a significant number from the raw image data. Instead, most of the fish during this hour appear to be small local fish, with occasional passage of individual sockeye. We will discuss this later.

Figure 17. shows the results for a period in which pink are expected to dominate. As can be seen, in the hours of high passage, the pink proportion is high (80% and up), whereas in the hours of low passage (from Hour 20 to 30), the sockeye proportion appears to be significant. Visual inspection indicates that this period of low fish passage is dominated by passage of individual fish, unlike the high passage period where fish typically come in groups. These individual fish include sturgeon, as well as sockeye and pink.

The averaged pink/sockeye composition over these periods can be calculated by separating the fish count in each dataset into pink count and sockeye count, based on the corresponding composition estimates. The total pink/sockeye count is obtained by summing the individual counts. The resulting averaged pink/sockeye composition is 0.27/0.73 for Fig. 16. (2007-08-19), and 0.79/0.21 for Fig. 17. (2007-08-25 to 26). This seems consistent with typical salmon migration patterns at Mission, in which majority of pink salmon arrive in late August and early September.

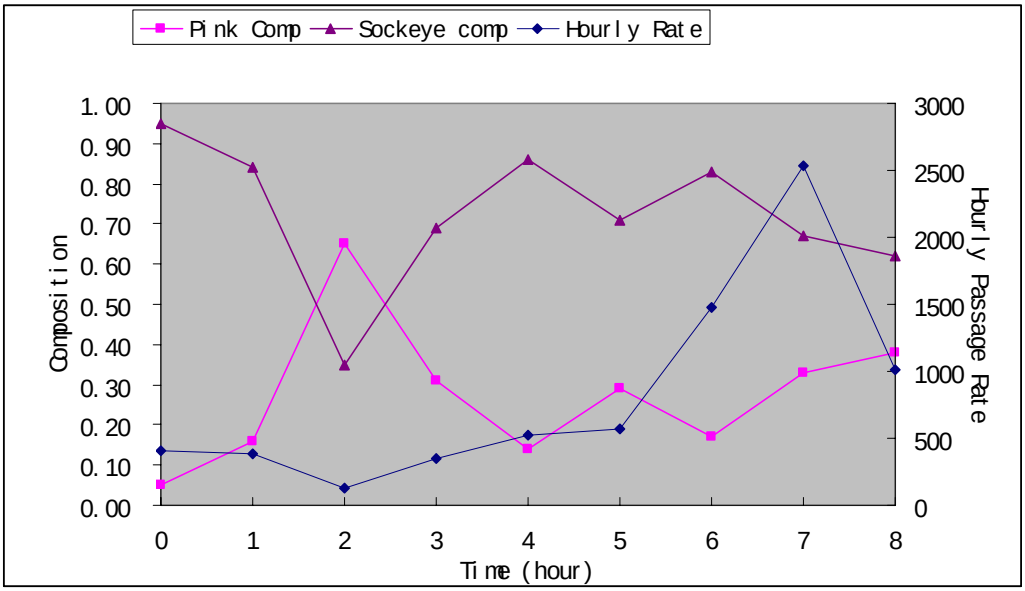

Fig. 16.: Pink/Sockeye composition estimates in Mission for the period Hour 0 – 8, August 19, 2007. The pink line represents the pink proportion and the brown one is for the sockeye proportion. The blue line is the total passage normalized to hourly rate. Each data point represents an estimate based on 20min raw data.

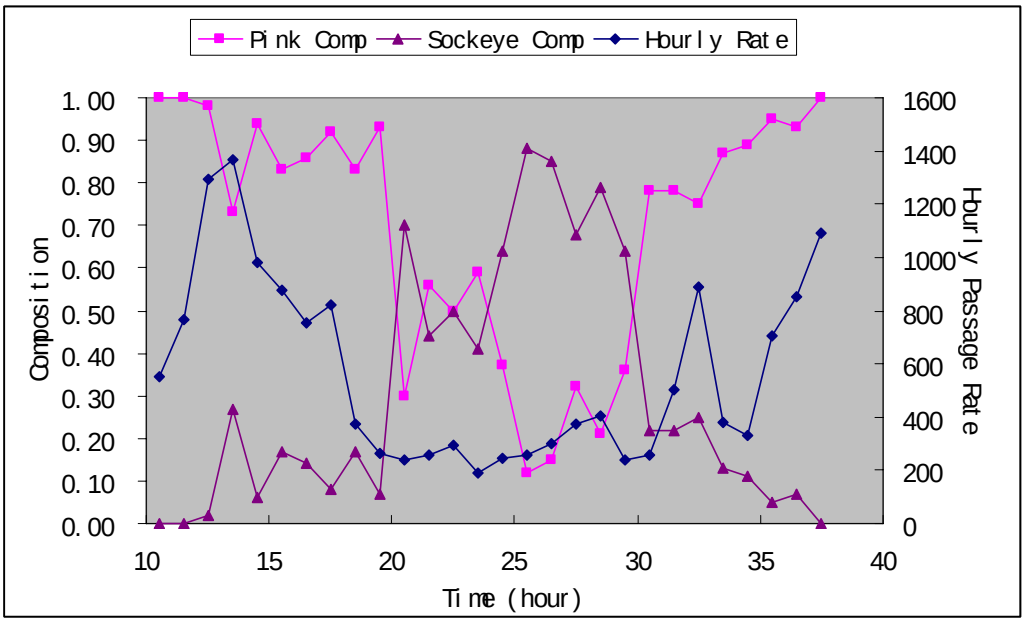

Fig. 17.: Pink/Sockeye composition estimates in Mission for the period from August 25, Hour 10, to August 26, Hour 13, 2007. The pink line represents the pink proportion and the brown one is for the sockeye proportion. The blue line is the total passage normalized to hourly rate. Each data point represents an estimate based on 25min raw data.

5.5 Discussion

In this work, we have investigated two species composition estimation methods, namely, the DFA classifier and the EM algorithm. We found that the DFA method was more reliable. The performances of the EM algorithm did not match those of the DFA classifier, as indicated by Table 1. (We have also investigated Neural Networks, and found that although the method yielded similar misclassification rates to the DFA method, its testing results were unstable. Given its computational cost, we did not continue to pursue Neural Networks.)

However, the DFA method requires good monospecific samples for training and testing, which may be difficult to obtain from field operations when fish of multiple species co-migrate (we actually picked training samples manually from a number of different datasets). Also, training data may have to be re-collected for different sites and in different system configurations, for the DFA classifier to achieve optimal results. From this perspective, the EM algorithm has some advantages, because it does not require good training data due to its iterative optimization procedure (it is possible to run the EM algorithm without training data). However, more research is needed to improve its inferior performances compared to DFA.

Table 1 indicates that the standard deviation of the DFA estimates of species proportions is 6-8%. This may be a problem when the proportions are close (e.g. 0.45/0.55). Although we accept the fact that there is always some variability in the estimates, the variability could be reduced by using ensemble methods, such as Bagging or Boosting (Han and Kamber, 2005). These methods create a set of bootstrap samples for training, leading to a set of classifiers. Input data are presented to these classifiers, and a decision regarding which class an input is assigned to, is based on the classification results from all the classifiers. In this way, the overall performances are often improved.

In the above analysis, we have assumed that there are only two species (pink and sockeye) in the data. However, field data often contain more species than assumed in the estimation model, and in this case, the estimation will be biased. As mentioned above, there are very few identifiable pink salmon at Hour 2 in Fig.16., despite an estimate of pink proportion at 70%. Most of the fish during that period appear to be local small fish. Because the length of pink is typically smaller than that of sockeye, those local fish are most likely to be classified as pink. It is also possible that sturgeon are partially responsible for the high sockeye proportion at Hour 25 in Fig. 17.. In addition, if field data contain non-fish targets such as debris, an additional class will need to be introduced to represent this type of targets.

The uncertainty in the number of classes in measurement data may be one of the major sources of bias in composition estimation. For more robust estimation of species composition, it may be necessary to estimate the number of distinct classes first, and then apply a classification method such as DFA to estimate species composition. Alternatively, we can also explore estimation methods that allow the number of species to be estimated directly from measurement data. For example, some unsupervised clustering algorithms, such as hierarchical algorithms, are able to estimate the number of clusters based on a certain criterion. Another approach is competitive learning (Theodoridis and Koutroumbas, 2003), which allows the number of classes to be determined while performing classification. However, when the number of species is known precisely and single species training data can be obtained (for example, via careful visual inspection), we expect the present estimation method (DFA) to perform at least as well as the unsupervised methods. The performances of various approaches can be assessed via numerical simulation as described in this report.

6 CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

In this project, we have conducted a feasibility study to determine whether DIDSON data collected at Mission can be used to estimate the composition of pink and sockeye in the field. We have performed an analysis of morphometric and behavioral characteristics of individual fish extracted from the DIDSON data, and selected a set of feature variables as shown in Fig. 12., which we think would have effective discrimination power. We have investigated three species composition estimation methods, and found that the DFA method was most reliable. The performances of the DFA classifier were evaluated via numerical simulation. It is found that even though the error rate of the classifier is not small (about 20%), the averaged bias of the composition estimates is 1% - 3%, because the error rate is balanced between the two species.

We have applied the DFA classifier to the 2007 DIDSON data. The results show that the composition can be highly variable from hour to hour, revealing dynamic passage patterns of different species. However, the averaged composition estimates over a period of several hours seem consistent with typical salmon migration patterns at Mission. In summary, our work has clearly demonstrated that it is entirely feasible to estimate the pink/sockeye composition based on DIDSON data.

6.2 Recommendations

Despite the success in this project, further research is needed to improve the accuracy of the estimation. As discussed in the previous section, there are several areas for improvement, which are summarized as follows:

- *Training data*: We have mentioned that training data are an essential part of the DFA method or other similar methods. In this project, we selected training samples from data collected in similar system configurations and deployment environments, and applied the DFA method to other data collected in such conditions. We need to conduct a systematic study on how training samples collected from different deployment sites and under different configurations would affect the estimation accuracy. This can only be possible as we build more databases over years.
- *Improved accuracy*: The simulation work shows that the DFA estimates have a certain degree of variability. One way to reduce the variability and improve accuracy is to use ensemble methods.
- *Estimation of number of classes*: One of the major sources of bias in composition estimation is the uncertainty in the number of classes in measurement data. We should explore some unsupervised clustering algorithms, such as hierarchical algorithms (Theodoridis and Koutroumbas, 2003), that have the ability to estimate the number of clusters based on a certain criterion.
- *Alternative algorithms*: One alternative approach is based on mixture models and uses Bayesian inference to estimate species composition from measurement data, similar to the aforementioned EM algorithm. It uses a Monte Carlo simulation method (Gibbs sampler) to assess the uncertainty in estimates (Gelman et al, 2004). The required number of components in the model is then determined by testing whether the resulting model adequately describes the observed data. Bayesian approaches have also been used in fisheries stock assessment (Meyer and Millar, 1999) and species composition estimation based on split-beam data (Fleischman and Burwen, 2003).
- *Implementation*: The aforementioned estimation methods are often mathematically complex and computationally intensive. However, recent advance in computational algorithms has made the approaches very practical. In this project, we have implemented the algorithms on a high-level language platform (e.g. MATLAB), which allows quick implementation but has limited flexibility

and performances, thus not suitable for practical use. Therefore, in addition to further research, it is necessary to build a stand-alone software framework for species composition estimation, providing users with flexibility to choose a specific estimation scheme suitable for their data environments.

7 ACKNOWLEDGEMENTS

We have collaborated with the Stock Monitoring Group at the Pacific Salmon Commission (PSC) in this project, who have provided the DIDSON data collected at Mission from the 2004-2007 management seasons, and valuable suggestions at various stages of this project.

8 REFERENCES

Belcher, E., W. Hanot, and J. Burch, 2002: *Dual-Frequency Identification Sonar, Proc. 2002 International Symposium on Underwater Technology*, Tokyo, Japan, pp. 187-192.

Blackman, S. S., and R. Popoli, 1999: *Design and Analysis of Modern Tracking Systems*. Artech House Publishers, Dedham, MA.

Cooke, R. C. and G. E. Lord, 1978: *Identification of Stocks of Bristol Bay Sockeye Salmon, OncorHynchus Nerka, by Evaluating Scale Patterns with a Polynomial Discriminant Method*. Fishery Bulletin, Vol. 76, No. 2, pp. 415-423.

Fleischman, S. and D. Burwen, 2003: *Mixture Models for the Species Apportionment of Hydroacoustic Data, with Echo-Envelope Length as the Discriminatory Variable*. ICES J. Mar. Sci., 60, pp592-598.

Gelman, A, J. B. Carlin, H. S. Stern and D. B. Rubin, 2004: *Bayesian Data Analysis, 2nd edition*, Chapman and Hall.

Gonzalez, R. C. and R. E. Woods, 2002: *Digital Image Processing, 2nd edition*, Prentice Hall.

Han, J. and M. Kamber, 2005: *Data Mining: Concepts and Techniques, 2nd edition.* Morgan Kaufmann.

Haralabous, J. and S. Georgakarakos, 1996: *Artificial Neural Networks as a Tool for Species Identification of Fish Schools*. ICES J. Mar. Sci., 53, pp. 173-180.

Holmes, J. A., G. Cronite, and H. J. Enzenhofer, 2005: *Feasibility of Deploying a Dual Frequency Identification Sonar (DIDSON) System to Examine Salmon Spawning Ground Escapement in Major Tributary Systems of the Fraser River, British Columbia*. Canadian Technical Report of Fisheries and Aquatic Sciences 2592.

Horne, J. K., 2000: *Acoustic Approaches to Remote Species Identification: a Review*. Fisheries Oceanography, 9:4, pp. 356-371.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York.

Meyer, R. and R. B. Millar, 1999: *Bayesian Stock Assessment using a State-Space Implementation of the Delay Difference Model.* Can. J. Fish. Aquat. Sci., Vol. 56, pp37-52.

Samarasinghe, S. (2006): *Neural Networks for Applied Sciences and Engineering*, Auerbach Publications.

Theodoridis, S. and K. Koutroumbas, 2003: *Pattern Recognition, 2nd edition*, Academic Press.

Xie, Y., A. P. Gray, F. J. Martens, J. L. Boffey, and J. Cave, 2005: *Use of Dual Frequency Identification Sonar to Verify Split-Beam Estimates of Salmon Flux and to Examine Fish Behaviour in the Fraser River*. Pacific Salmon Commission Tech. Rep. No 16.

9 APPENDICES

9.1 Composition Estimation Based on Discriminant Function Analysis

DFA is based on the assumption of normality and Bayesian Decision theory (see McLachlan, 1992, for more details). That is, the probability distribution of samples is assumed to be a multivariate normal distribution. The parameters of the normal distribution are then estimated from training samples. Classification is based on Bayes Decision Theory, which classifies a sample into Group ω_i, if the posterior (*a posteriori*) probability for this group is the maximum. The posterior probability is the probability that an unknown sample belongs to a particular class, given its feature variables associated with the sample. It can be derived from the likelihood probability based on the Bayes rule:

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})}$$
$$p(\mathbf{x}) = \sum_{i=1}^{G} p(\mathbf{x} \mid \omega_i)P(\omega_i) \qquad \textbf{(1)}$$

where $P(\omega_i)$ is the prior (*a priori*) probability of a data sample belonging to Group ω_i, $p(\mathbf{x}|\omega_i)$ is the probability distribution of samples from Group ω_i (likelihood function) and $p(\mathbf{x})$ is the probability distribution of the data. If the prior probability is known or can be estimated from training samples, classification can be performed based on the maximum of the posterior probabilities.

However, in an application where group proportions (which determine prior probabilities) are to be estimated based on classification, careful consideration should be given to prior probabilities. Unless there is clear distinction between groups, the choice of prior probabilities has a significant effect on classification. This poses a dilemma. Cook and Lord (1978) suggest a method in which prior probabilities are chosen to balance misclassification rates in test data. That is, the proportion of samples from one group assigned to other groups is balanced by the samples from other groups assigned to this group. The chosen prior probabilities are fixed for subsequent applications.

9.2 Composition Estimation Based on Expectation Maximization

The Expectation Maximization (EM) algorithm is suited for cases where the available data set is incomplete. For example, a set of samples is known to have been drawn from G groups, but we lack the information regarding which sample is drawn from which group. We can, however, define a complete data set $y_k = (\mathbf{x_k}, j_k)$, where j_k indicates that the k-th sample is drawn from the j-th group. Then we have

$$p(\mathbf{x}_k, j_k; \boldsymbol{\theta}) = p(\mathbf{x}_k \mid j_k; \boldsymbol{\theta})P_{j_k},$$

where $\boldsymbol{\theta}$ is a set of unknown parameters in the probability distribution, and P_j is the prior probability which is also unknown. Assuming mutual independence among the samples, we can define a log likelihood function as

$$L(\boldsymbol{\theta}) = \sum_{k=1}^{N} \ln(p(\mathbf{x}_k \mid j_k; \boldsymbol{\theta}) P_{j_k}) .$$

The EM algorithm first takes the expectation of the likelihood function (*E-Step*) over the unobserved data (j_k). That is,

$$\begin{aligned} Q(\boldsymbol{\theta}, \mathbf{P}) &= E\left\{ \sum_{k=1}^{N} \ln(p(\mathbf{x}_k \mid j_k; \boldsymbol{\theta}) P_{j_k}) \right\} = \sum_{k=1}^{N} E\left\{ \ln(p(\mathbf{x}_k \mid j_k; \boldsymbol{\theta}) P_{j_k}) \right\} \\ &= \sum_{k=1}^{N} \sum_{j_k=1}^{G} P(j_k \mid \mathbf{x}_k; \boldsymbol{\Theta}) \ln(p(\mathbf{x}_k \mid j_k; \boldsymbol{\theta}) P_{j_k}) \end{aligned}$$

where $\boldsymbol{\Theta}=(\boldsymbol{\theta}, \mathbf{P})$ represents all the unknown parameters to be found. Since for each k, we sum up over all possible j and these are all the same for all k, we can simply drop the subscript k from j. The above equation becomes

$$Q(\boldsymbol{\Theta}) = \sum_{k=1}^{N} \sum_{j=1}^{G} P(j \mid \mathbf{x}_k; \boldsymbol{\Theta}) \ln(p(\mathbf{x}_k \mid j; \boldsymbol{\theta}) P_j) .$$

The posterior probability can be calculated based on Bayes' rule:

$$P(j \mid \mathbf{x}_k; \boldsymbol{\Theta}) = \frac{p(\mathbf{x}_k \mid j; \boldsymbol{\theta}) P_j}{p(\mathbf{x}_k;)}$$

$$p(\mathbf{x}_k; \boldsymbol{\Theta}) = \sum_{j=1}^{G} p(\mathbf{x}_k \mid j; \boldsymbol{\Theta}) P_j$$

Then the algorithm maximizes (*M-Step*) the above function with respect to $\boldsymbol{\Theta}$. However, this step will lead to a set of nonlinear equations, which can be solved via an iterative approach.