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 2010 SOUTHERN FUND
PROJECT CLOSURE REPORT

In-Season Estimation of Salmon Species Composition using DIDSON Sonar Image Data Collected at Mission

PREPARED FOR
PACIFIC SALMON COMMISSION

JANUARY, 2011

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

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. In a previous project funded by the Southern Fund (2007-2008), we have been able to extract fish morphometric and behavioral information from DIDSON image data, and apply the Discriminant Function Analysis (DFA) algorithm to classify each individual fish and obtain an estimate of species composition. The current project is a follow-up aimed at streamlining the data processing in our estimation approach and making it practical for field applications. In this project, we have converted the algorithms developed in the previous project from MATLAB code into a user-friendly, standalone software system, streamlining the estimation procedure and incorporating various statistical and computational tools. We have also used the software to process all the available 2007 DIDSON data collected in the Fraser River at Mission and provided by the Stock Monitoring Group at the Pacific Salmon Commission (PSC). The results are compared with test fishing data.

We have analyzed 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 a suitable estimator in terms of performance and stability. The performances of the DFA method were evaluated via numerical simulation, and it was found that even though the error rate may be up to 20%, the averaged bias of the composition estimates is less than 1% after compensation for the error rate. The results of using the DFA method to estimate salmon composition (sockeye v.s. pink) for the 2007 data show a temporal pattern consistent with typical salmon migration patterns at Mission, in which the majority of pink salmon arrive in late August and early September. Other independent estimates from test fishing and fish wheel also show a similar pattern. Our DIDSON estimates appear to be in the same range as the test fishing estimates, but are significantly higher than the fish wheel results. However, our estimates also offer detailed dynamic passage patterns of different species, which are not available with other techniques. 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. Further research is also recommended.

2 INTRODUCTION

In a previous project funded by the Southern Fund (2007-2008), entitled 'A feasibility study of using DIDSON imaging sonar to estimate species composition at Mission' (Vitech 2008), we have been able to extract fish morphometric and behavioral information from DIDSON image data, and apply the Discriminant Function Analysis (DFA) algorithm to classify each individual fish and obtain an estimate of species composition. The current project is a follow-up aimed at streamlining the data processing in our estimation approach and making it practical for field applications. This is because the original estimation approach involves complex mathematical computations, and was developed on the MATLAB platform, which is not practical for field applications. In this project, we have converted the MATLAB code into a user-friendly, standalone software system, streamlining the estimation procedure we developed and incorporating various statistical and computational tools. We have also used the software to process all the available 2007 DIDSON data collected in the Fraser River at Mission. The results are compared with test fishing data.

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 (IntelliHAT) 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 are described in the previous project report (Vitech 2008).

To perform species identification, we need to construct a set of feature variables (descriptors) from morphometric and behavioral characteristics derived from DIDSON data using IntelliHAT. 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.

3.2 Derivation of Fish Characteristics from DIDSON Data

Here we describe the characteristics (referred to feature variables) of individual fish we derived from DIDSON data for species classification. As described in the 2007 report, raw DIDSON

data are first processed by IntelliHAT, which outputs track data that contain the position, target size (automatically measured on the image), and acoustic intensity of each target tracked over frames. Feature variables are then derived from these track data. We have identified the following variables as most relevant:

- 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.
- Cross-beam velocity (speed and direction), which provides an indicator of the upstream/downstream direction for migrating salmon.
- Aspect ratio of the imaged fish, which 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.
- Fish length, which was discussed in detail in the 2007 report. Since the measured length of an imaged fish during its lifetime can be highly variable, the fish length is defined as the value obtained when the 'brightness' of the imaged fish is maximum.
- Brightness, which is defined as the total acoustic intensity in an 'imaged' fish, normalized by the number of the pixels associated with the fish. This parameter is measured at the same time as fish length. 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.
- Local density, which measures the fish density in the neighborhood of a fish.

3.3 Optimal Estimation of Live Fish Length

In our previous report (Vitech 2008), we have described an automatic approach to measurement of live fish length in the field, which will be described here briefly for convenience. First, for each tracked fish on an image, we use a simple automatic thresholding algorithm (Gonzalez and Woods, 2002) to determine automatically an intensity threshold in an area surrounding the fish. Pixels in this area with intensity above the threshold are considered to be due to the fish. The fish length is then calculated from the selected pixels. Then the fish length is measured over frames for the tracked fish, leading to a time series of length for each tracked fish. The time series typically shows great variability, for a number of reasons: 1) fish sometimes swim in tight aggregations, making it difficult to measure unambiguously; 2) fish often flex as they move across the acoustic view, making their visibility greatly variable over time; 3) fish at some locations are less visible due to background interference. 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.

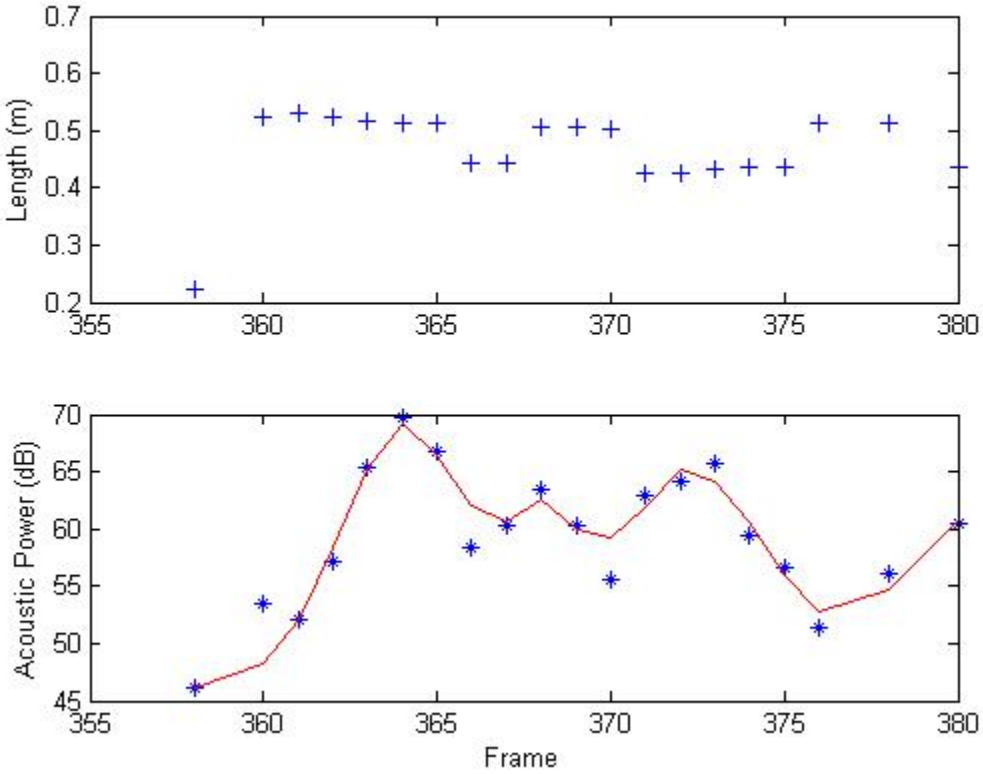

Fig. 1.: Length measurements of a live fish as a function of time (frames). The lower plot is the corresponding 'brightness' of the fish image (see the text for more explanation).

Figure 1 (upper plot) shows an example of the length time series of a fish as it swims across the acoustic beam, where the horizontal axis is frame number. Now the question is which measurement represents the best estimate of the fish length, given the variability. Intuitively, it is reasonable to assume that the length measurement is most reliable when the fish image is brightest. The brightness (see the brightness definition in Section 3.3) corresponding to the length measurements is shown in the lower plot of Fig. 1. 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.

3.4 Calculation of Local Density

Local density is the target density in the neighbor of a fish. This is also a dynamic parameter varying as the fish moves. It is calculated in the following procedure:

1. Track a video file and generate a corresponding track file.
2. For each target (fish) in the track file, examine each frame in which the fish appears. Use an unsupervised clustering algorithm (reference??) to find natural clusters in the target positions on the current frame. Calculate the maximum spatial extent in each cluster, and define local density as the number of targets in the cluster divided by the spatial extent.
3. Repeat the calculation for each frame where the target exists. Find the maximum value to represent the local density of the target.

3.5 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 is model-based, where species composition is derived directly from a set of observations based on a model, without having to classify individual observations. In this project, we focus on the methods in the first group.

W employed two classification-based methods, one of which is Discriminant Function Analysis (DFA), 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 a classifier 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 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 **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.

3.6 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.

3.7 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 Software Suite for Species Composition Estimation

4.1 FACE (Fish Analysis & Composition Estimation)

The software (FACE) we built for this project implements the estimation procedure described in the preceding section. It has a simple main interface as shown in Figure 2, and consists of four modules: feature variable generation, classifier training, performance evaluation, and application. These will be described in the following subsections. The software and a more detailed user manual have been released to the Pacific Salmon Commission (Contact Dr. Yunbo Xie).

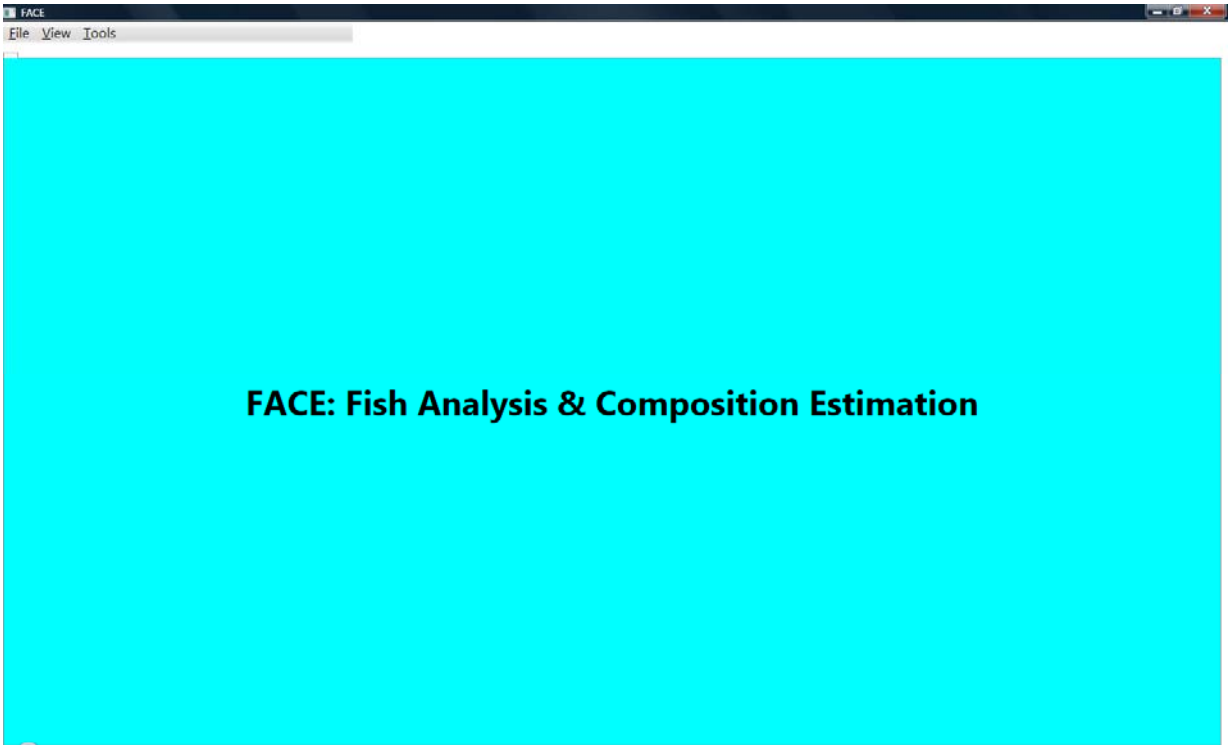

Fig. 2.: Interface of the software FACE.

4.2 Generating Feature Variables

The first step is to extract feature variables from track data files generated by Vitech's software IntelliHAT (see Section 3). The feature variable generation module is shown in Fig.3, where the input and output files are shown in the text boxes on the right. Four plots on the left show the histograms of swim speed, fish length, brightness (acoustic power), and local density, for each individual output file (the highlighted one in the output box). Files output from this module are used in the subsequent training, evaluation, and application.

Fig. 3.: Feature variable Generation Module

4.3 Analyzing Feature Variables

This module allows the user to performance more in-depth analysis and visualization of feature variable data (work in progress). At this moment, it only facilitates visualization of distributions of feature variables. As can be seen in Fig. 4, the module can show distributions in individual files by selecting the file name in the list box on the right. It also allows the user to group files according to their classes (e.g. training data), and shows distributions of all data in one class.

Fig. 4.: Feature Variable Analysis module.

4.4 Training a Classifier

The training module is shown in Fig. 5. The user first selects feature variable files for each class (species), and then selects a classifier to train (e.g. DFA or Random Forests). After training is finished, the classification matrix will be shown in the box on the right. Also, for DFA, the log-likelihood for each species will be shown on the plot on the left. For Random Forests, the variable importance is also shown in the box in the middle.

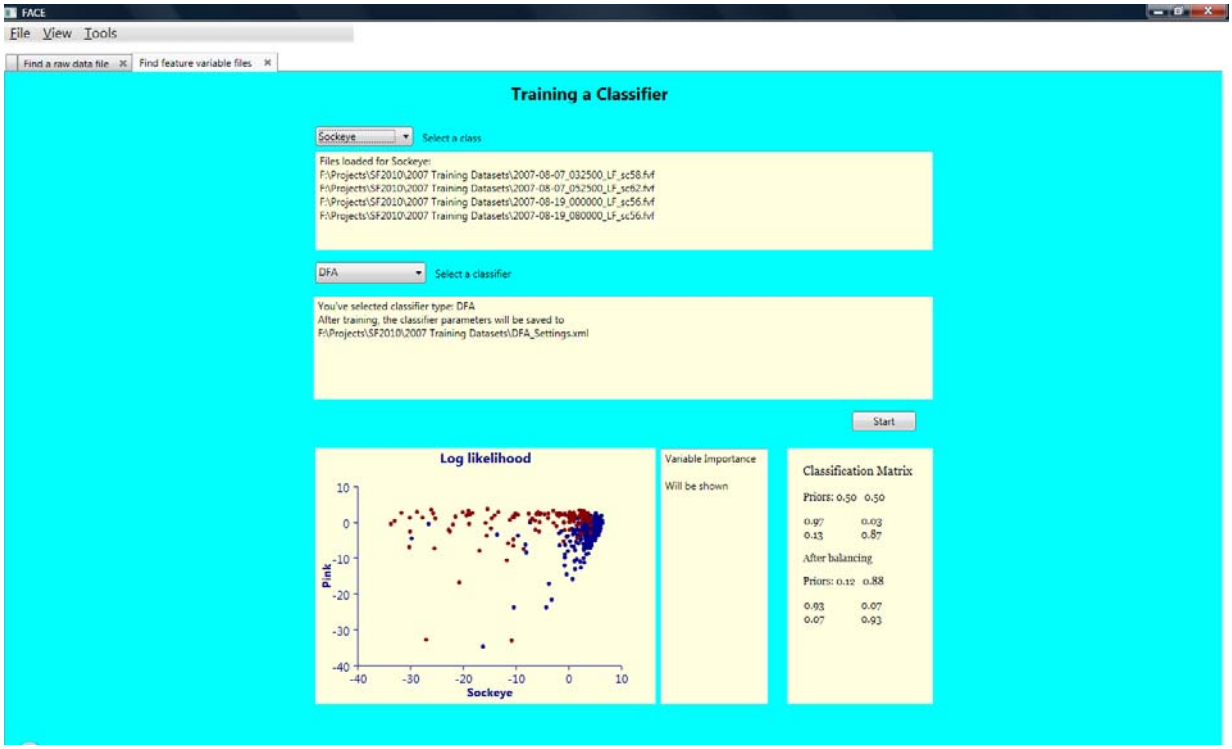

Fig. 5.: Training module.

4.5 Evaluating a Classifier

For a set of training data, we can perform a more extensive training and performance evaluation, via numerical simulation. This involves resampling a subset from the training data with replacement for training, and using the remaining data points for performance evaluation (see Section 3 for more details). The simulation can be run repeatedly several times, and a different subset of data is obtained at each run. Figure 6 shows the evaluation module, where the user can load training data files (on the left), select a classifier, and enter the number of simulation runs. After the simulation runs are finished, the distribution of classification error rates will be displayed as a histogram (in the middle). In addition, given a trained classifier, the user can also evaluate the performance of its use in composition estimation (on the right). In this case, the user can set an assumed sockeye proportion, and the software will generate samples with the sockeye proportion equal to the prescribed one, and use the classifier to classify each sample and thus estimate the composition.

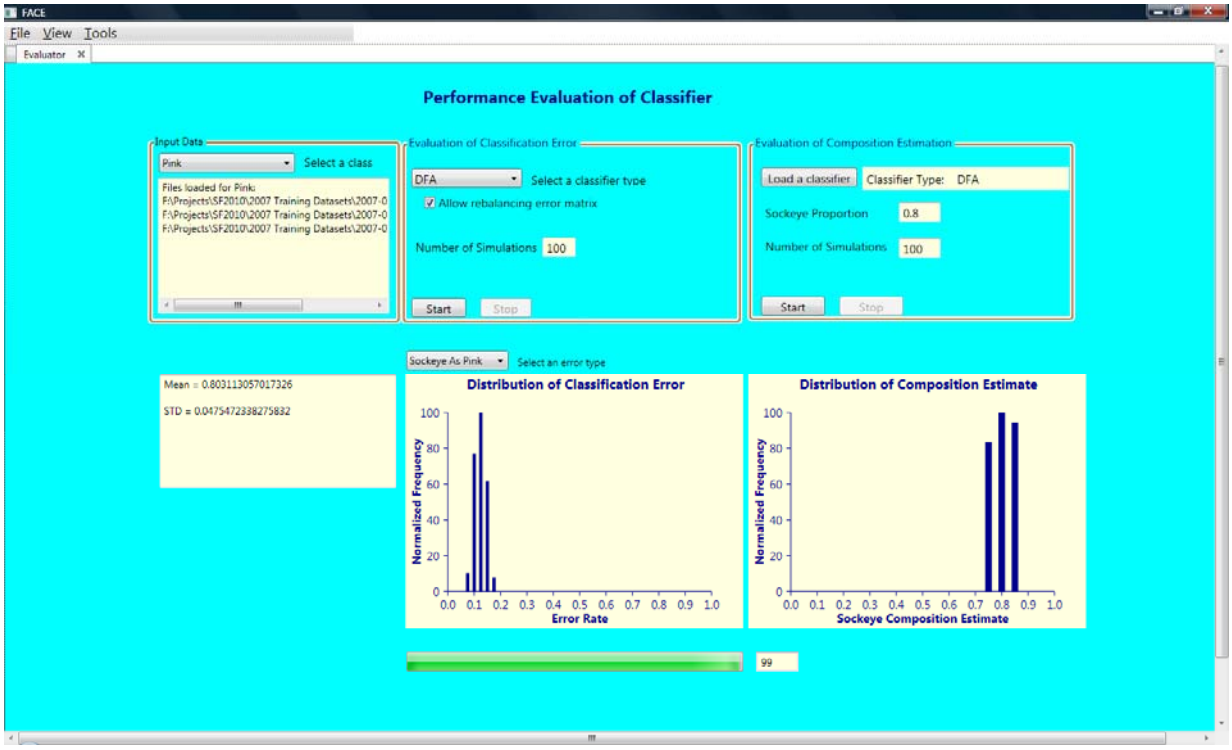

Fig. 6.: Evaluation module

4.6 Applying a Classifier

Finally, after a classifier is trained and its performance is deemed satisfactory, the user can apply it to new data in the application module shown in Fig. 7. The user first loads a parameter file for the classifier, and then a set of new data files for classification. On the output, composition estimates for each data file are provided, with the classification error of the classifier compensated for. The result can also be exported to a CSV file which can be opened by MS Excel for further analysis.

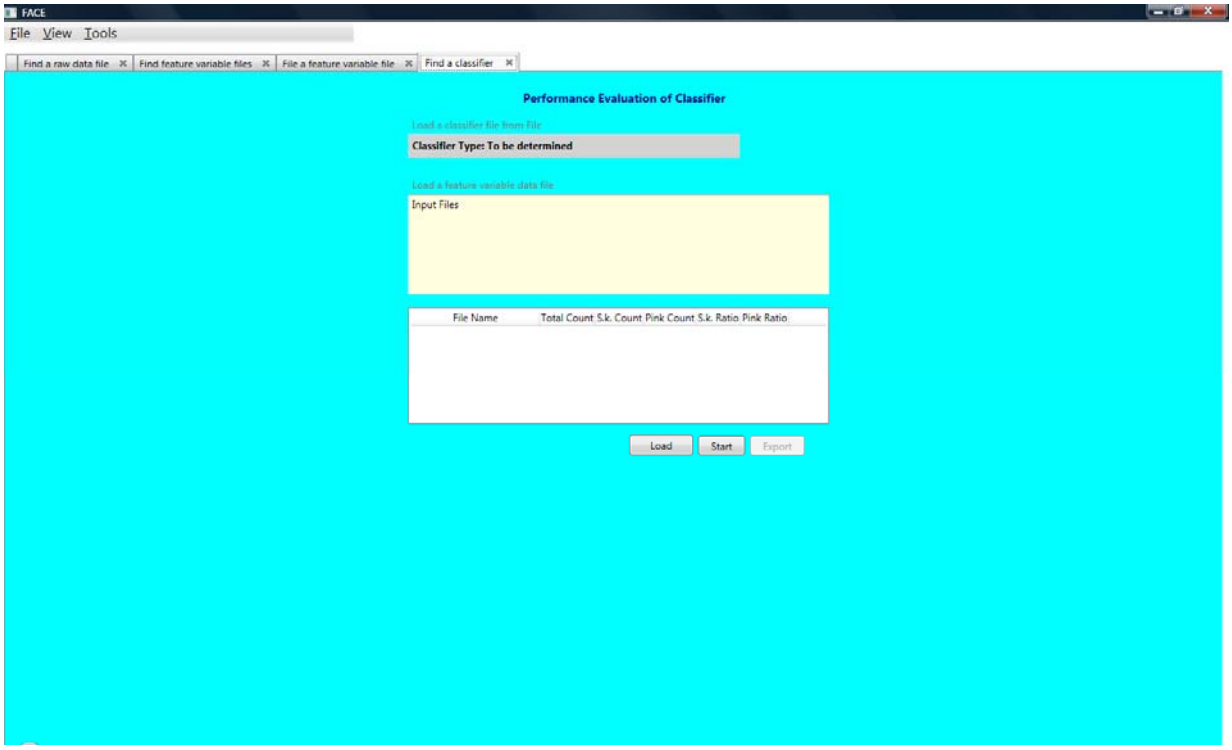

Fig. 7.: Application module.

5 RESULTS and DISCUSSION

5.1 Training and Evaluation of the DFA Classifier

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. Figures 8 and 9 show histograms of some feature variables derived from monospecific data (Sockeye and Pink).

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 10-20cm in length on average, their distributions of length are expected to overlap as seen in the length histograms in Fig. 8 and 9, where it can be seen that the sockeye length histogram peaks at 50-60cm, and pink length peaks peaks at 30-40cm. Another feature that seems to be better able to distinguish the two species is acoustic backscatter intensity (brightness as defined in Section 3), although it is correlated with fish length. As can be seen in Fig. 8 and 9, the pink data have a broader distribution with the peak around 60 dB, while the sockeye data have a somewhat narrower distribution peaked around 68 dB. 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 in Fig. 8 and 9 show that pink 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. The narrower speed distribution of pink is probably due to the grouping behavior

of pink, which we may characterize with a parameter such as local target density (see Section 3). As can be seen from the distributions in Fig. 8 and 9, the local density of sockeye is essentially less than 1.0, which the pink density is spread over 0 to 10.0, with a significant number of data over 10.0. So these two distributions appear to be very noticeably different.

Another parameter we also use is cross-beam swim speed, which is 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.

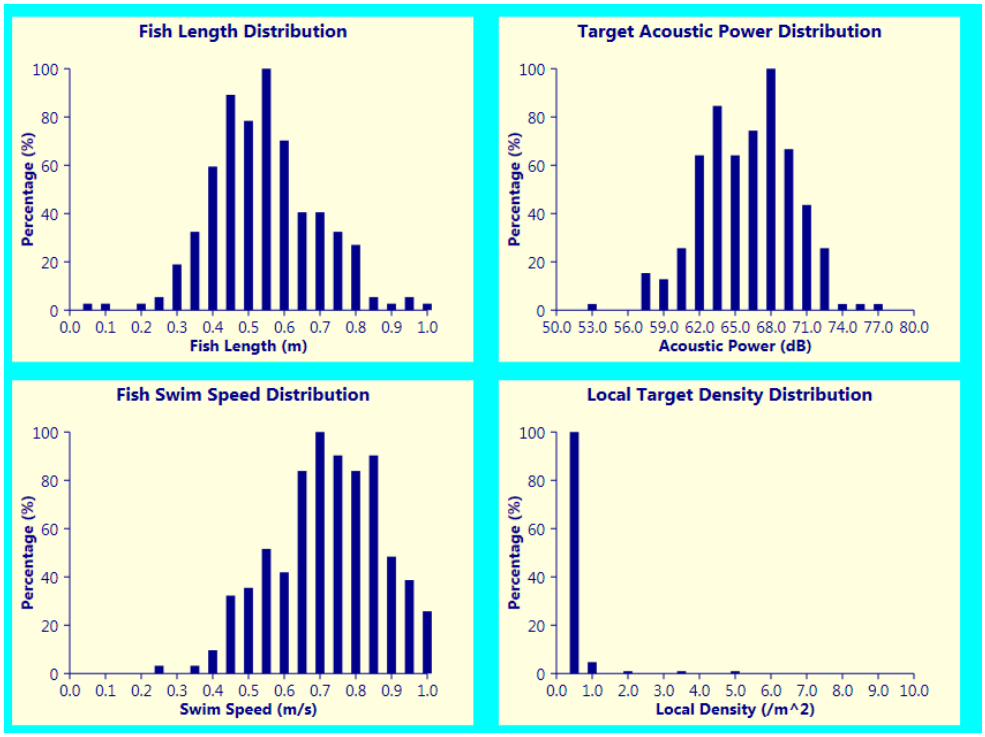

Fig. 8.: Histograms of feature variables of the monospecific training data set (Sockeye). Top-left: fish length; Top-right: Acoustic Power; Bottom-left: Swim Speed; Bottom-right: Local density.

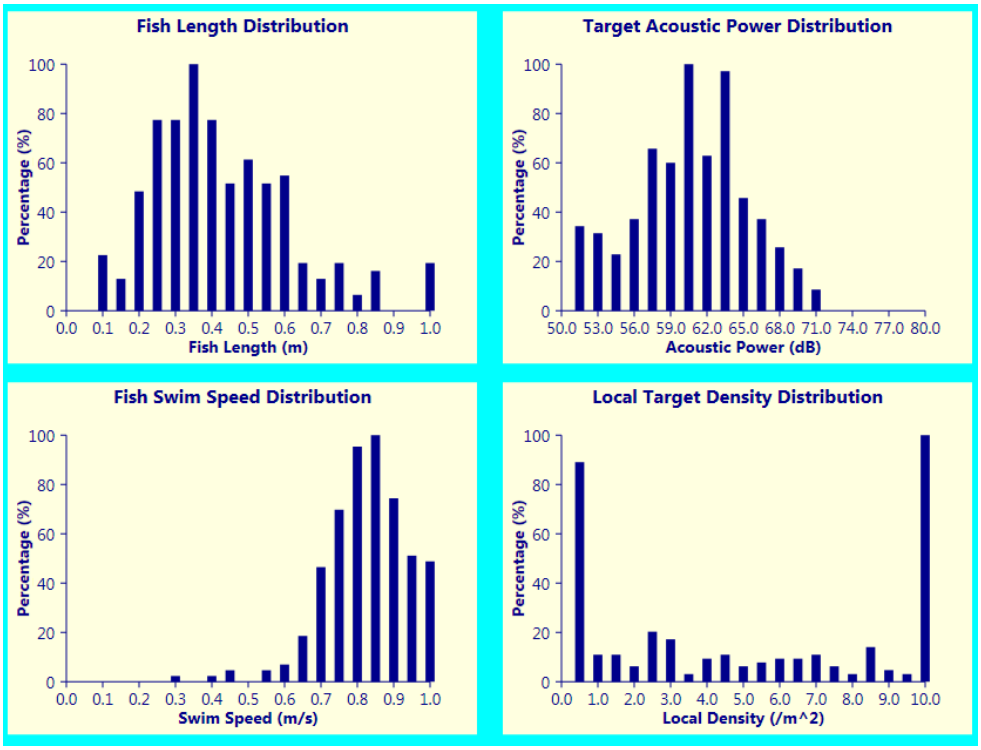

Fig. 9.: Histograms of feature variables of the monospecific training data set (Pink). Top-left: fish length; Top-right: Acoustic Power; Bottom-left: Swim Speed; Bottom-right: Local density.

We used the above four parameters plus the velocity component in the cross beam speed as feature variables to train a DFA classifier. We then used the same training to get a quick evaluation of the classifier. This led to a classification error 3% (sockeye misclassified as pink) and 22% (pink misclassified as sockeye), when the prior probabilities are set equal. This indicates an asymmetric discriminating power of the classifier. However, after choosing the prior probabilities to balance the classification errors (Section 3), the error rate becomes 12% (see the table below).

Classification Matrix

Priors: 0.50 0.50

0.97	0.03
0.22	0.78

After balancing

Priors: 0.14 0.86

0.88	0.12
0.12	0.88

Table 1: Classification matrix of the DFA classifier.

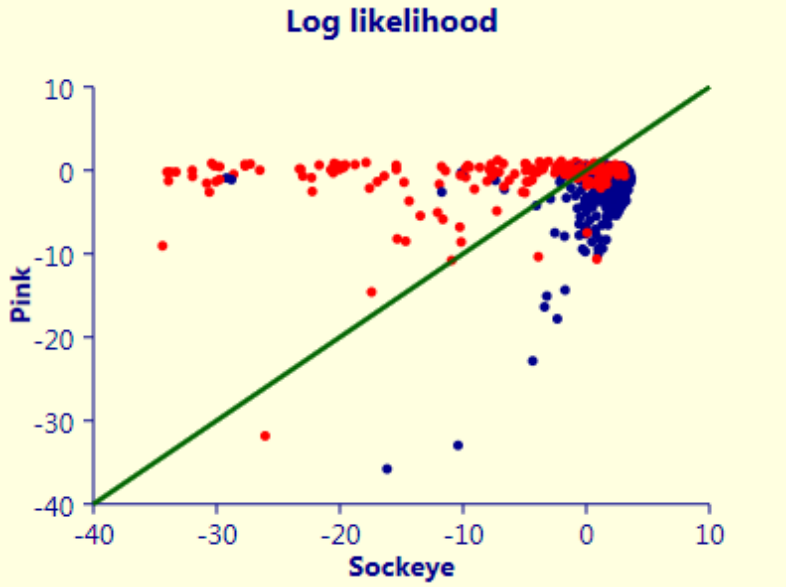

Fig. 10. Histogram of log-likelihood ratio calculated by the DFA classifier, for the training data set. The upper pane is the result obtained with pure pink samples as input, and the lower one for pure sockeye samples.

The distributions in Fig.8 and 9 show considerable overlap, although those for sockeye are generally narrower. However, when these data are fed to a classifier, such as the DFA classifier, the classification enjoys a reasonably high accuracy. We may gain further understanding on the discriminating power of the selected feature variables by calculating the likelihood (posterior probability) of an input data vector being classified to a certain class (see Section 3). Figure 10 shows a scatter plot of the likelihood (in log-scale) for the trained DFA, given the monospecific training data (sockeye and pink), in the case of equal prior probabilities. The horizontal (or vertical) axis is the likelihood of being classified as sockeye (or pink). The red dots are the likelihoods of the pink training data, while the blue ones are from the sockeye training data. The line indicates the boundary above which a data point will be classified as pink based on the DFA principle (assuming equal prior probabilities). It can be seen in Fig. 10 that most the blue dots fall in the lower region, which indicates that sockeye are very unlikely to be misclassified as pink. However, there are a substantial amount of red dots located *below* the decision line, indicating pink are somewhat likely to be misclassified as sockeye. This is why the classification error shows asymmetry before rebalancing (see the matrix in the table). After balancing, the DFA classifier offers a much better separation between the two species than what the feature variable distributions appear to suggest.

5.2 Error Analysis of Species Composition Estimation

Table 1 shows the performance of the trained DFA classifier based on the training data, which are not independent testing data. To perform a more objective evaluation, we use the bootstrap method described Section 3 to generate separate sets of training and testing data. The simulation can be performed over many realizations of training and testing data, leading to a distribution of the rate of misclassification, as shown Fig. 11, where the simulation was performed 500 times with the prior probabilities chosen to balance the rate of misclassification, leading to a mean error of 12.1% with standard deviation of 2.2%.

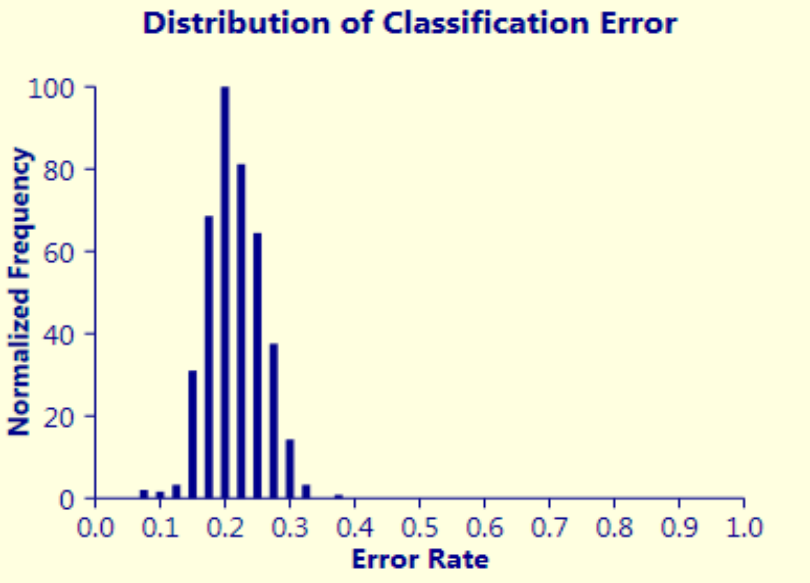

Fig. 11.: Histogram of classification error rate by the DFA classifier. The mean is 0.121 and the standard deviation is 0.022.

We can also use the same approach to evaluate the ability of a classifier to estimate species compositions of pink and sockeye. That is, we randomly take a different number of samples separately from the training data of sockeye and pink, to simulate different compositions. For each composition, the simulation is performed a number of times (say 500), and this provides us with a distribution of the composition estimate. Table 2 shows the results of such simulations for the DFA estimates of different sockeye proportions. It is seen that even though the error rate of the classifier is not small (12%), the bias of the composition estimates is quite small, because the error rate is balanced between the two species and compensated for. The standard deviations of these estimates range from 2.7% to 3.7%. It is also noticed that the standard deviation is higher when the proportion of one species is very small, than when the proportions are close

True Pink Prop	Mean DFA Est.	STD DFA Est.
0.1	0.09	0.033
0.2	0.19	0.036
0.3	0.30	0.034
0.4	0.40	0.027
0.5	0.51	0.028
0.6	0.60	0.028
0.7	0.70	0.037
0.8	0.80	0.029
0.9	0.90	0.036

Table 2: Performance of DFA estimates of different sockeye proportions.

5.3 Application of the DFA Classifier to Field Data

Having evaluated the DFA classifier, we now use it to estimate species composition sockeye and pink composition in the time period of August 6-30, 2007 (based on all available DIDSON data available in this period). We chose data collected under system configurations similar to those for the training data, where 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. Figure 12 shows the DIDSON estimates of sockeye proportion (shown as big blue circles) during this period, where test fishing (TF) estimates and fish wheel (FW) estimates of the composition are also shown in for comparison. The TF data (shown as blue diamonds) before Aug 15 are from in-river TF, whereas after Aug 15, they are projected from marine area TF (courtesy of PSC). In both DIDSON and TF estimates, only two species (sockeye and pink) were considered, while the fish wheel data (shown as red diamonds, provided by LGL Limited) include a third class. All the estimates show a transition period from sockeye dominance to pink dominance from Aug 10 to 25, but the DIDSON and FW estimates show a more consistent transition while the TF estimates projected from marine area show the sockeye composition back upwards from Aug 22 to 25. On the other hand, the FW sockeye estimates appear to be significant less than the DIDSON and TF estimates (there is a third class in the FW, whose proportion is not significant enough to explain the difference).

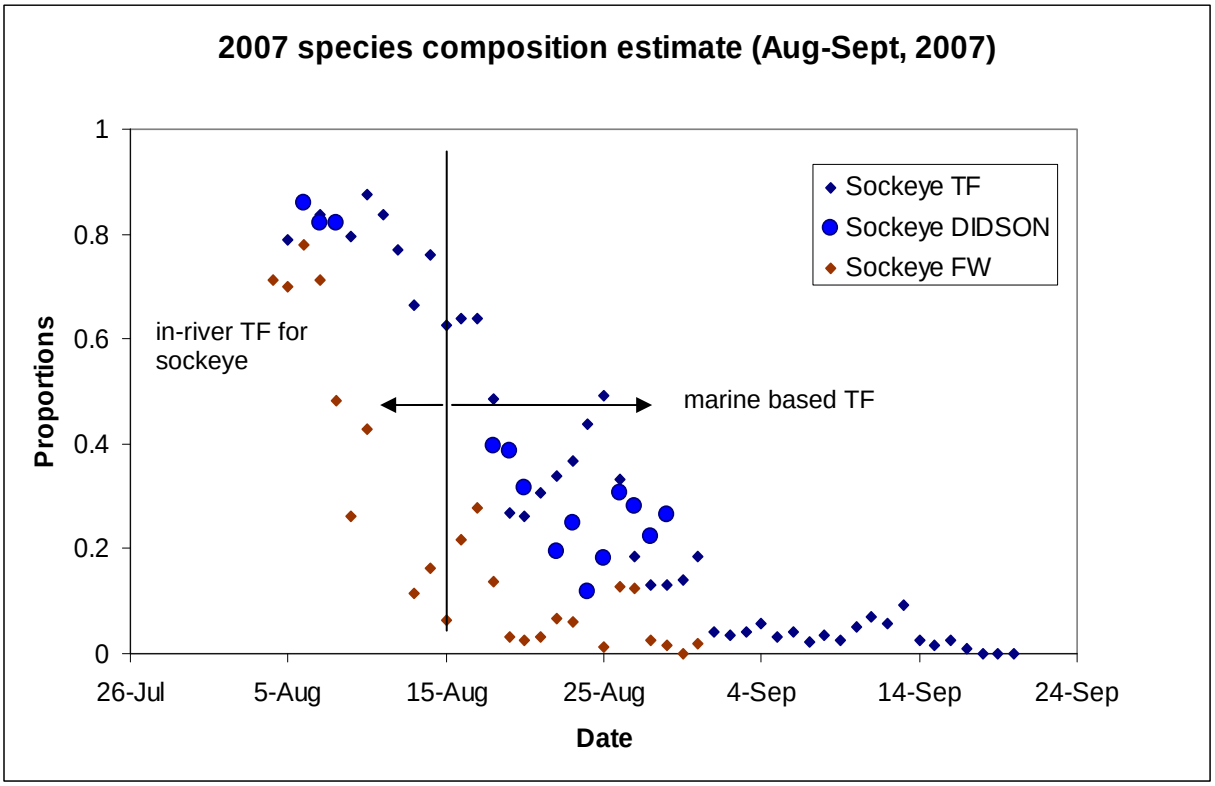

Fig. 12.: Sockeye composition estimates in Mission in August and September of 2007. Pink squares and blue diamonds represent the pink proportion and the sockeye proportion respectively, estimated from test fishing (TF) data. Before Aug 15, the TF estimates are derived from in-river test fishing, whereas after Aug 15, they are projected from marine area TF results. Big blue circles and pink triangles are the DIDSON estimates.

Figure 13 shows hourly variability of the DIDSON estimates from Aug 17 (hour 01) to Aug 20 (hour 17), 2007, together with fish passage rate, which has been normalized to hourly rate for comparison (note that the passage rate is for reference only, due to track filtering to reduce noise. See the discussion below). As can be seen, the composition can be highly variable from hour to hour, revealing dynamic patterns of passage of different species. In a single day, for example, the sockeye composition can vary from around 10-20% in the hours of high passage, to 90-100% in the hours of low passage. Visual inspection also confirms that the periods of low fish passage are dominated by passage of individual fish, where in the high passage periods fish typically come in groups. Thus, if data collection by non-acoustic methods is conducted only a few hours a day, there is potential for a large bias in composition estimates. One big advantage of our sonar-based approach is that it can provide continuous monitoring and estimates.

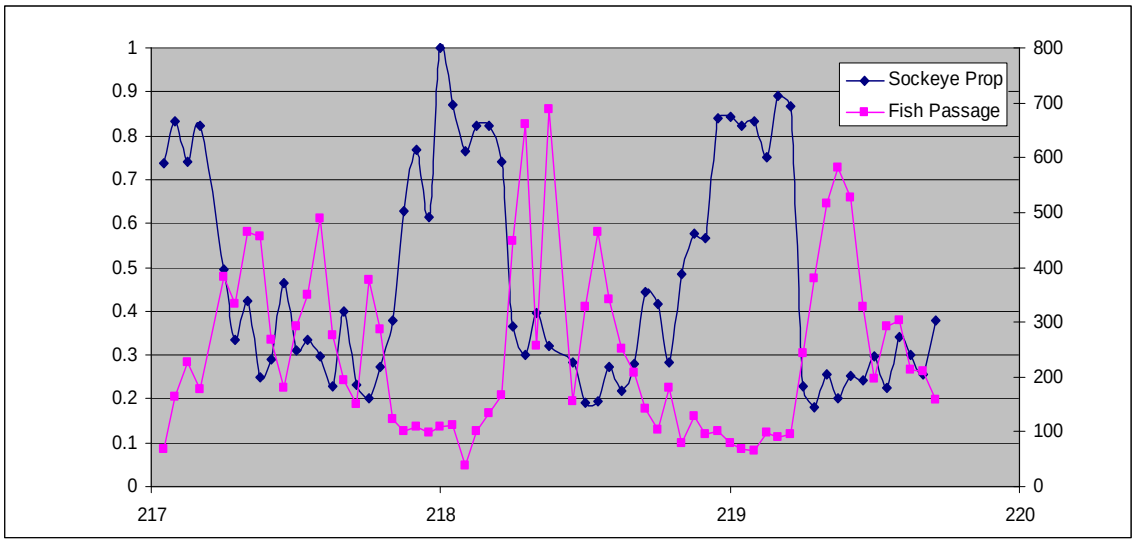

Fig. 13.: Time variability of sockeye composition (blue curve, left axis), together with fish passage (pink curve, right axis) for the period of August 18 (hour 01) to August 20 (hour 17), 2007. The horizontal axis is Julian day.

5.4 Challenges

The results presented here were not obtained without challenges, the most significant of which is interference of noise. This may result from bubbles, debris, and milling or holding fish, but also from unstable image background due to shifting in sonar orientation. An example is shown in Fig. 14. We have used a filter to reduce the noise. As described in Section 3, we first tracked moving targets in the video data and generated corresponding track files. This filter calculates a set of fish behavior parameters from on the track data, and eliminates tracks whose parameters fail to pass certain criteria. However, in the case of high density noise as in Fig. 14, there will be always residual noise after filtering, which sometimes can be rather significant. We currently manually clean the residual noise. After the filtering and cleaning, we allowed the software to generate feature variable data for every remaining track. More research is needed to develop an efficient filter to minimize manual efforts and assess the effects of filtering on final results. One approach is to treat the noise as a different class than pink and sockeye.

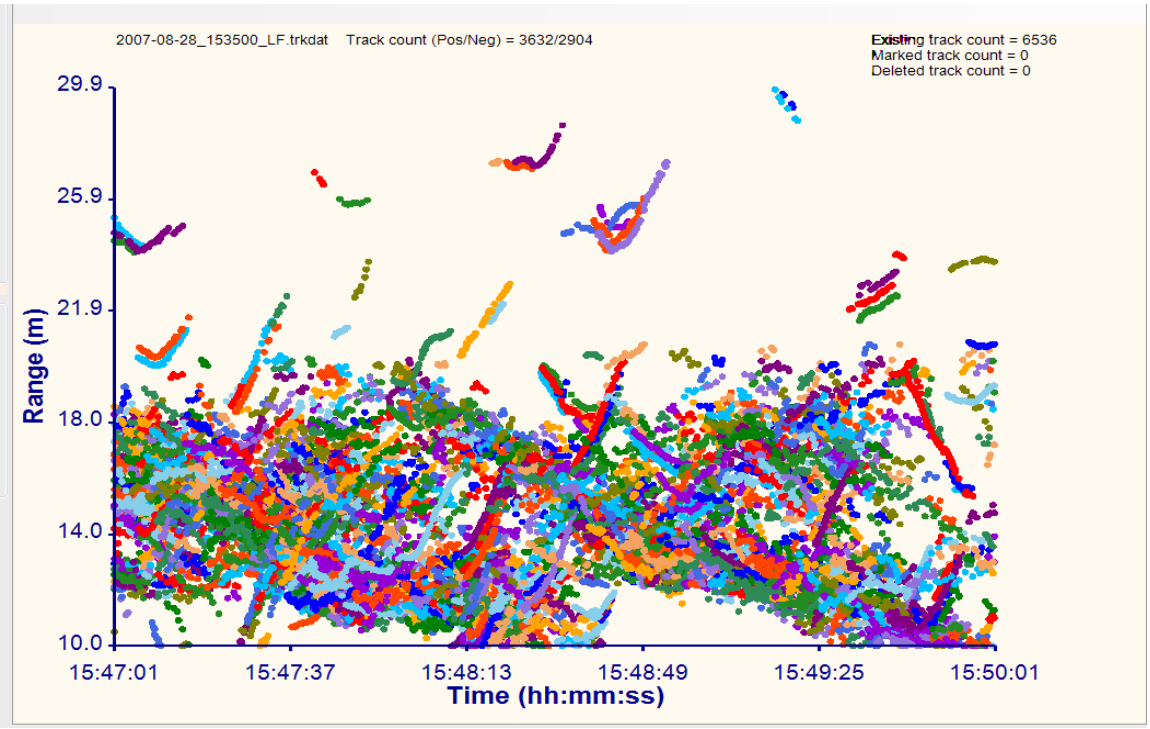

Fig. 14.: Heavy noise interfering with tracking results. Individual tracks can be clearly identified for range greater than 20m, but are embedded in noise clouds in range under 20m.

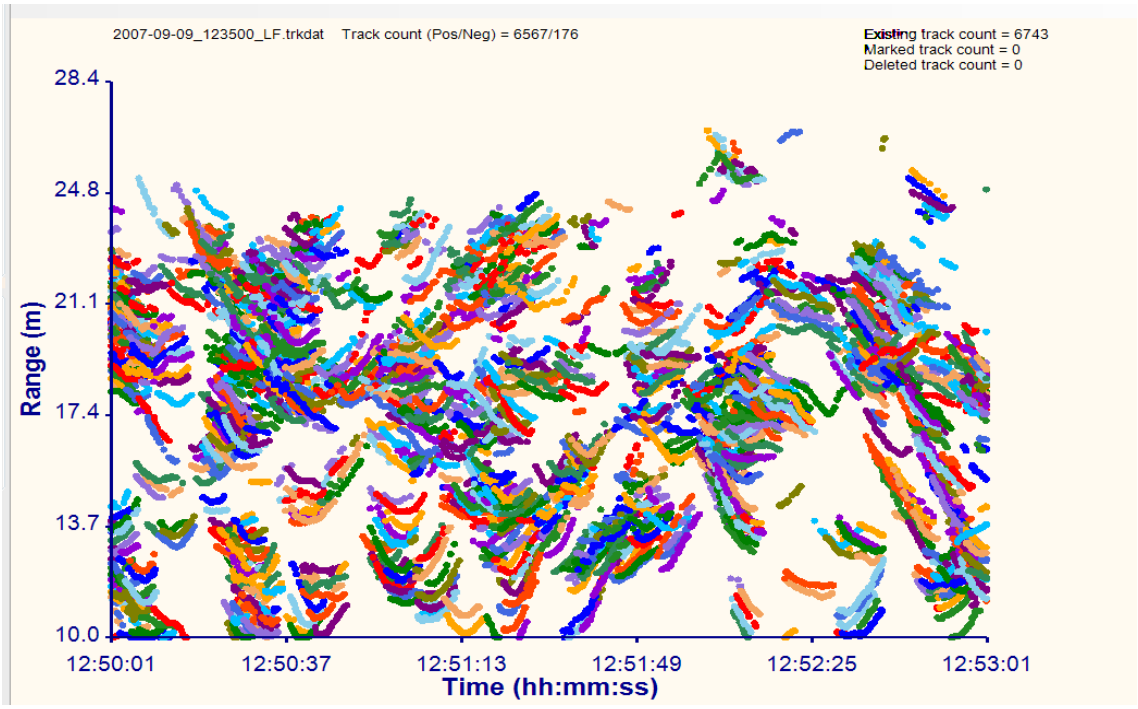

Fig. 15.: Extremely high passage of pink causes trouble for our current approach.

The second challenge is high fish passage: our current approach becomes less effective when fish passage increases. For example, pink passed the sonar in high volume in early September of 2007, with a passage rate of up to 30,000 per hour (see Fig. 15 for an example). Our approach has difficulty handling this kind of data. More work is needed to find a remedy or determine an upper limit of fish passage for DIDSON-based composition estimates.

The third challenge is interference of local resident fish: these are typically smaller than pink and sockeye, and concentrated in the near range from the sonar. In the above analysis, we have assumed that there are only two species (pink and sockeye) in the data. When there are more species than assumed in the estimation model, the estimation will be biased. For example, those local fish are more likely to be classified as pink since the length of pink is typically smaller than that of sockeye. Figure 16(a) shows a fish length distribution in which we can see three identifiable peaks, with the peak around 20cm consistent with the length of local fish.

To minimize the effect, we have applied a range gate to filter out the local fish since in many cases they were spatially separable from salmon. However, a more preferable approach is to introduce a third class (rather than pink and sockeye) if there is significant spatial overlap between local fish and salmon. 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. 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. Such an approach also has an additional benefit that allows separation of local fish from sockeye in the early run of sockeye.

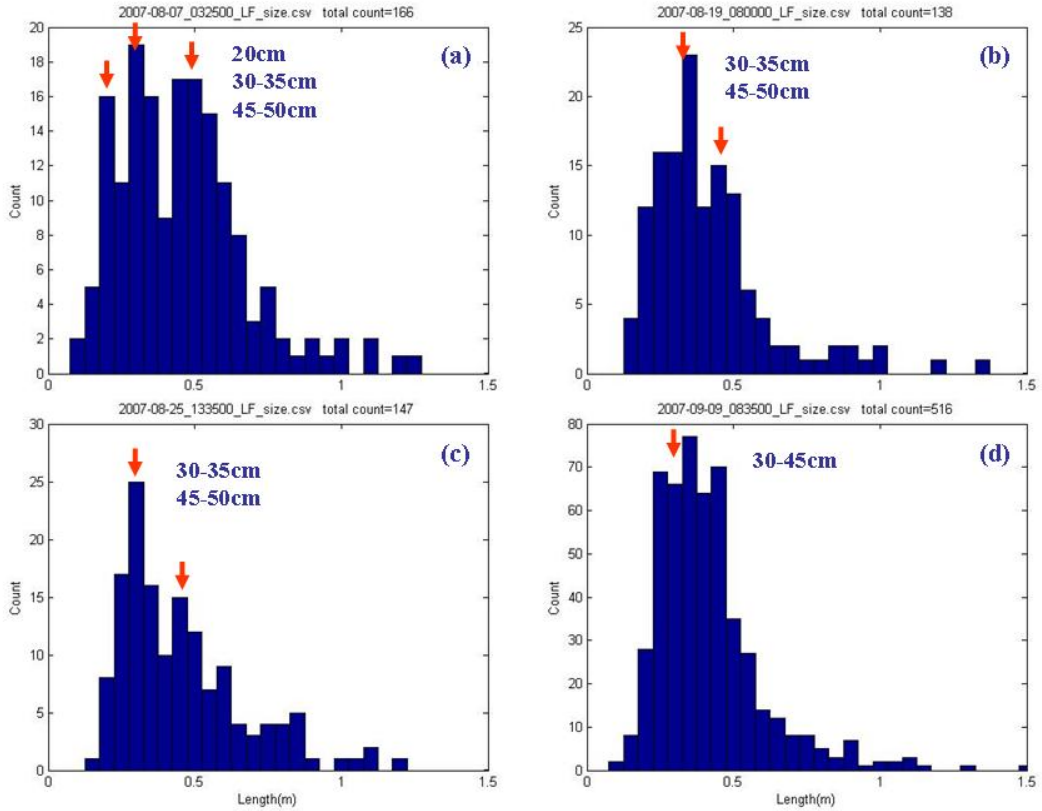

Fig. 16.: Histograms of length measurements from 2007 Mission DIDSON data collected on four days. Each data set has a time span of 20min. (a) 2007-08-07; (b) 2007-08-19; (c) 2007-08-25; (d) 2007-9-09. Note that the peak at 45-50cm in plot (a), (b) and (c) gradually diminishes, and that it disappears 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.

6 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:

- *Interference Noise*: Here we refer to noise as any interference that is manifested in images and subsequent processing. This may result from bubbles, debris, and milling or holding fish, but also from unstable image background due to shifting in sonar orientation. Although we have used a filter to reduce the noise, there will be always residual noise after filtering, which sometimes can be rather significant. We currently manually clean the residual noise. However, manual cleaning significantly hinder our capability of automatic processing. More research is needed to develop an efficient filter to minimize manual efforts and assess the effects of filtering on final results.

- *Estimation of number of classes*: Given the existence of local resident fish in the early run of sockeye, which will inevitably affect the estimate of pink/sockeye composition, it is desirable to be able to estimate their proportion in the total fish count. A suitable approach is to introduce a third class (rather than pink and sockeye) in the model. 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. 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. Some unsupervised clustering algorithms, such as hierarchical algorithms (Theodoridis and Koutroumbas, 2003) have the ability to estimate the number of clusters based on a certain criterion. Development of such an approach also has an additional benefit in the period when pink do not return and the pink/sockeye composition is not required, since by allowing automatic separation of local fish from sockeye, it helps to reduce human labor needed to perform manual estimates of local fish.
- *High fish passage*: High fish passage occurs when pink return in large amounts. Our current approach becomes less effective when fish passage rate increases to 10,000-30,000 per hour. More work is needed to find a remedy or determine an upper limit of fish passage for DIDSON-based composition estimates

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

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.

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.

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

Lu, H. J. and K. T. Lee, 2005: Species identification of fish shoals from echograms by an echo-signal image processing system. *Fisheries Research*. Volume 24, Issue 2, Pages 99-111

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

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

Vitech Innovative Research and Consulting, 2008: A feasibility study of using DIDSON imaging sonar to estimate species composition at Mission. SEF 2007 Report submitted to Pacific Salmon Commission.

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.