Copy No. _____

Defence Research and Development Canada
Recherche et développement pour la défense Canada

DEFENCE R&D DÉFENSE

Receiver-Operating-Characteristic (ROC) analysis applied to listening-test data

Measures of performance in aural classification of sonar echoes

Nancy Allen

Defence R&D Canada – Atlantic

Technical Memorandum
DRDC Atlantic TM 2007-353
August 2008

Canada

This page intentionally left blank.

Receiver-Operating-Characteristic (ROC) analysis applied to listening-test data

Measures of performance in aural classification of sonar echoes

Nancy Allen

Defence R&D Canada – Atlantic

Technical Memorandum
DRDC Atlantic TM 2007-353
August 2008

Principal Author

Original signed by N. Allen

N. Allen

Defence Scientist

Approved by

Original signed by N. Sponagle

N. Sponagle

Head/Underwater Sensing

Approved for release by

Original signed by R. Kuwahara for

J. L. Kennedy

Head/Document Review Panel

This report was produced for DRDC project 11cq11.

In conducting the research described in this report, the investigators adhered to the policies and procedures set out in the Tri-Council Policy Statement: Ethical conduct for research involving humans, National Council on Ethics in Human Research, Ottawa, 1998 as issued jointly by the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada and the Social Sciences and Humanities Research Council of Canada.

Abstract

A Receiver-Operating-Characteristic analysis was conducted on data from two listening tests that were carried out as part of a DRDC Technology Investment Fund project on aural classification of sonar echoes. All the human listeners were DND personnel with significant experience in sonar. An automatic classifier was also tested. The results of the analysis strongly support the idea of using aural cues to discriminate between target echoes and clutter. The automatic classifier outperformed some of the human listeners and was on par with the others. Performances of the human listeners and automatic classifier in the second listening test, where a high-pass filter was applied to the echoes to remove all frequency content below 500 Hz, were definitely poorer than for the first test, where the full available bandwidth of 0 □2 kHz was exploited, but still substantially better than chance performance.

Résumé

On a proc□d□ □une analyse de la fonction d'efficacit□ du r□cepteur (ROC) portant sur les donn□es de deux essais d□□coute effectu□s dans le cadre d□un projet du Fonds d'investissement technologique de RDDC concernant la classification auditive des □chos sonar. Tous les □couteurs humains participants □taient des employ□s du MDN ayant une exp□rience appr□ciable du sonar. Un classificateur automatique a □galement □t□ mis □ l□essai. Les r□sultats de cette analyse appuient fortement le projet d□utiliser des indices auditifs pour faire la distinction entre les □chos de cibles r□elles et le clutter. Le classificateur automatique a surclass□ certains des □couteurs humains et s□est av□r□ l□□gal des autres. Lors du deuxi□me essai d□□coute o□ l□on a appliqu□ un filtre passe-haut □ l□ensemble des □chos pour □liminer les □l□ments sonores d□une fr□quence inf□rieure □500 Hz, la performance des □couteurs humains et celle du classificateur automatique ont □t□ manifestement moins bonnes que lors des premiers tests o□ l□on avait exploit□ toute la largeur de bande disponible (0 - 2 kHz), mais elle fut tout de m□me de beaucoup sup□rieure au simple hasard.

This page intentionally left blank.

Executive summary

Receiver-Operating-Characteristic (ROC) analysis applied to listening-test data: Measures of performance in aural classification of sonar echoes

Allen, N.; DRDC Atlantic TM 2007-353; Defence R&D Canada – Atlantic; August 2008.

Background: Human-performance testing was one component of a DRDC Technology Investment Fund project on aural classification of sonar echoes. Two listening tests were designed, approved by the DRDC Human Research Ethics Committee, and carried out at DRDC Atlantic. The sample echoes that served as stimuli for the tests were obtained from recordings of an active broadband sonar experiment and also served as test cases for developing an automatic classifier. The main difference in the tests was the bandwidth of the stimuli. In the first listening test, the full 2-kHz bandwidth that was available was used. A high-pass-filtered version of the same stimuli was used in the second test, which removed any frequency content below 500 Hz. The volunteers for the tests were military and civilian DND personnel with significant experience in sonar. Their main task in the tests was a rating exercise. During this exercise they would listen to a sequence of echoes and, for each echo, decide if it was a target echo or clutter and assign a level of confidence to the decision. Their responses provided the input data for the Receiver-Operating-Characteristic (ROC) analysis that was carried out to model individual and average performance.

Results: Statistical models known as binormal ROC curves were fitted to the human-performance data. The measure of performance A_z, which corresponds to the area under the binormal ROC curve, was used to quantify the performances of individual listeners and groups of listeners. All results were significantly higher than the 0.5 index that represents chance performance. It was also obvious that performance in classifying the full-bandwidth echoes was significantly better than performance in classifying the reduced-bandwidth echoes. Results from an automatic classifier were also analyzed in a similar way to the human-performance data. A statistical test proved that the automatic classifier outperformed at least two of the 13 listeners in the full-band test and five of nine listeners in the reduced-band test; it was on par with the other listeners. Additional quantitative and qualitative analyses were carried to out to explain why a small proportion of the human-performance data that were collected could not be adequately modelled.

Significance: The results of this analysis strongly support the idea of using aural cues to discriminate between target echoes and clutter. The results also support the potential usefulness of the automatic aural classifier in operational systems: its performance was on par with the better human listeners. Unlike some of the human listeners who reported that fatigue was hindering their usual performance, the automatic classifier was most consistent in its results.

Future plans: There are plans to include the automatic aural classifier in a sonar demonstration system and to explore its applicability in other areas of sonar.

Sommaire

Receiver-Operating-Characteristic (ROC) analysis applied to listening-test data: Measures of performance in aural classification of sonar echoes

Allen, N.; DRDC Atlantic TM 2007-353; R & D pour la défence Canada – Atlantique; August 2008.

Contexte : La mise à l'épreuve des performances humaines était l'une des composantes d'un projet du Fonds d'investissement technologique de RDDC portant sur la classification auditive des échos sonar. Deux tests d'écoute ont été mis au point; ils ont ensuite été approuvés par le Comité d'éthique en matière d'étude sur des sujets humains de RDDC, puis exécutés par RDDC Atlantique. Les échantillons d'échos ayant servi de stimulus pour ces essais ont été tirés d'enregistrements réalisés dans le cadre d'essais expérimentaux d'un sonar actif à large bande; ils ont aussi servi de jeu d'essai pour la mise au point d'un classificateur automatique. Le principal facteur de distinction de ces tests était la largeur de bande des stimulus. Le premier essai mettait en jeu toute la largeur de bande disponible, soit 2 kHz. Lors du deuxième essai, on a par contre utilisé un filtre passe-haut qui éliminait tous les éléments sonores inférieurs à 500 Hz. Les volontaires ayant pris part à ces essais étaient des employés civils et militaires du MDN ayant une expérience appréciable du sonar. Leur principale tâche consistait à faire un exercice de classification. Le test consistait à écouter une série d'échos et à décider, dans chaque cas, si l'on avait affaire à une cible réelle ou à du clutter, tout en cotant le degré de confiance associé à la décision. Les réponses fournies ont servi d'intrant lors de l'analyse de la fonction d'efficacité du récepteur (ROC) que l'on a menée pour modéliser les performances individuelles et la moyenne.

Résultats : Des modèles statistiques connus sous le nom de courbes ROC binormales ont été adaptés aux données de la performance humaine. La mesure de la performance A_z qui correspond à la plage sous la courbe ROC binormale a servi à quantifier les performances des écouteurs individuels et celles des groupes d'écouteurs. Tous les résultats ont été substantiellement plus élevés que l'index de 0,5 qui représente la performance aléatoire. Il était aussi manifeste que la performance de classification des échos faisant intervenir la pleine largeur de bande était de beaucoup supérieure à celle où l'on avait employé la largeur de bande réduite. Les résultats d'un classificateur automatique ont aussi été analysés d'une façon similaire à la méthode utilisée pour les humains. Un test statistique a démontré que le classificateur automatique surclassait son pendant humain dans au moins deux des 13 tests d'écoute utilisant la pleine largeur de bande et dans cinq des neuf situations où l'écouteur testait la largeur de bande réduite, les autres étant à égalité. Des analyses quantitatives et qualitatives supplémentaires ont été effectuées pour tenter d'expliquer pourquoi une petite quantité de données afférentes à la performance humaine qui ont été recueillies n'a pu être modélisée adéquatement.

Portée : Les résultats de cette analyse appuient fortement le projet d'utiliser des indices auditifs pour faire la distinction entre les échos de cibles réelles et le clutter. Ils sont aussi favorables à l'éventualité d'utiliser un classificateur auditif automatique dans des systèmes opérationnels, la performance de ce dernier ayant égalé celle des humains les plus doués. Cependant, contrairement

□certains d□entre eux qui ont signal□que la fatigue alt□rait leur performance, le classificateur automatique obtient des r□sultats tr□s uniformes.

Recherches futures : On envisage d□inclure le classificateur auditif automatique dans un syst□me sonar de d□monstration pour en tester l□applicabilit□□d□autres facettes du sonar.

This page intentionally left blank.

Table of contents

List of figures

List of tables

Acknowledgements

The author is very grateful to the DND personnel from ADAC(A), CFMWC, CFNOS, MOG5, PSU HALIFAX, TRINITY, and DRDC Atlantic who volunteered for the listening tests, and to their Commanding Officers for supporting their participation. Also gratefully acknowledged are Mr. Victor Young and Dr. Paul Hines of DRDC Atlantic for supplying results from their automatic classifier and for valuable discussions on human and machine performance.

1 Introduction

1.1 Listening tests for a project on aural classification of sonar echoes

A recent DRDC Technology Investment Fund (TIF) project investigated the use of aural features to discriminate between sonar echoes from target-like objects and those referred to as geoclutter. Background information and a work outline were presented in the TIF project proposal [1]. One question of interest to the project was *Can humans hear differences between target echoes and geoclutter, and if so, how well?* The two main reasons for the question were

- Anecdotal evidence from within the sonar-operator community had suggested that echoes from man-made objects and those from naturally occurring features were aurally distinguishable, but this evidence had not been scientifically verified; and
- Despite growing interest and incentive to develop automatic classifiers, there were no baseline measures for evaluating their performance.

One such automatic classifier, based on aural features associated with musical timbre discrimination, was developed during the course of the TIF project [2] [3]. The data used in its development and testing consisted of a set of echoes recorded during a sonar experiment involving explosive broadband sources. These echoes were also used as stimuli in two sets of human listening tests that were designed and conducted, during the same period, to address the above question. A detailed description of these human performance tests, including background information and rationale, was provided in the written protocols that were submitted and approved by the DRDC Human Research Ethics Committee [4]. The aspects of the tests that are relevant to this Receiver-Operating-Characteristic (ROC) analysis, which attempts to quantify the participants performance, will be reviewed in this report.

1.2 Performance of a binary classifier

A rating exercise was the main task in the tests. During this exercise, the participant would listen to a series of echoes and, for each echo, decide if it was a target echo or clutter and assign a level of confidence to the decision.

In choosing between target echo and clutter the listener behaved as a binary classifier or, equivalently, a detector.[1] Ideally, a detector will correctly identify all the target echoes and correctly reject all the clutter; in practice, it will miss some target echoes and mistake some clutter

[1] *Detection* and *binary classification* are used interchangeably in Signal Detection Theory to refer to the aim of identifying events of interest and rejecting events of no interest the latter being either noise or events that resemble events of interest but are not. The terms will also be used interchangeably here. But the reader should keep in mind that this work and the TIF project in general are concerned with aural distinctions in target echoes and false signals, i.e. clutter. In military operations one would usually refer to this challenge as one of *classification* only, since the term *detection* is reserved for the task of identifying signals from background noise.

for target echoes, i.e. produce false alarms. The correct responses and incorrect responses can be represented by four conditional probabilities related as follows:

Probability of detecting targets + *Probability of missing targets* = 1 (1)

Probability of correctly rejecting clutter + *Probability of producing false alarms* = 1 (2)

These probabilities are used to describe how a detector performs under particular operating conditions; this performance can be graphically represented as a point on a ROC graph (see Figure 1).

The detector can modify its performance to a certain extent to meet the demands of different operating environments, i.e. it has a range of operating points. The notion of a detection threshold is used in Signal Detection Theory for describing this flexibility. If for example the cost of missing a target is very high compared to the cost of producing false alarms, the detector can set a low threshold value to trigger at the slightest indication of an event of interest. The result is a high probability of detection combined with a relatively high probability of false alarm. In another situation, the detector may be discarded as unreliable if it produces too many false alarms. To limit the false-alarm rate, the detector may adopt the strategy of increasing its detection threshold, the disadvantage of course being an increased probability of missing targets. The example in Figure 1 shows a series of operating points for a detector, each corresponding to a particular threshold setting.

Also shown in Figure 1 is the ROC curve for that particular detector. It graphically depicts the inherent discrimination ability of a detector, often referred to as sensitivity[2]. The ROC curve in Figure 1 is re-displayed in Figure 2, together with two other sample curves. The ROC curve of the ideal detector is the point in the top left-hand corner of the graph (or, equivalently, two lines, one following the y-axis and the other following the top of the graph). The ROC curve of the detector that cannot discriminate at a better-than-chance level is the line along the positive major diagonal of the graph. The ROC curve of a "good" detector, such as the one shown, lies somewhere in between. All realistic ROC curves illustrate that an improvement in the probability of detection is achieved at the expense of an increase in the probability of false alarm, and vice-versa.

[2] The term *sensitivity* is sometimes limited to only describing the probability of detection (see, for example, the explicit distinction made between sensitivity and specificity made in Section 1.3.2 of [7]). It will be used here in its more general sense of inherent ability to discriminate between two classes of objects.

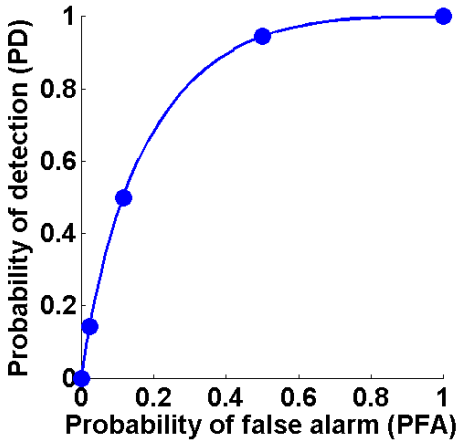

Figure 1: ROC graph showing a number of operating points for a given detector, which represent its performance under different operating conditions. The smooth curve is known as the ROC curve of the detector.

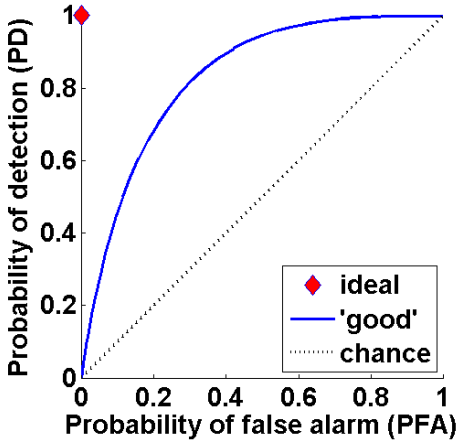

Figure 2: ROC curves depicting ideal performance, an example of "good" performance, and chance performance

The ROC curve is a graphical model of the detector's sensitivity; numeric indices of performance also exist [5][7]. Their relative conciseness and simplicity compared to graphical information is usually achieved at the expense of loss of information or a requirement to make one or more assumptions. Consider for example the performance index of area-under-the-curve, one of the easiest to grasp. The areas corresponding to the three curves shown in Figure 2 are 1 (ideal), 0.5 (chance), and approximately 0.8 ("good"). Although simple, this index does not provide information about the shape of the ROC curve (e.g. amount of symmetry about the negative major

diagonal of the ROC graph), which may be of interest if one wanted, for example, to identify the best detector to use for specific operating conditions.

One recognized technique for measuring a detector's performance at a number of operating points, *n* say (and ultimately for obtaining a representative curve of performance) is to construct a basic YES/NO decision test (YES for target echo and NO for clutter in the case of the listening tests) using a single set of stimuli and to repeat the test *n* times, each time imposing different operating conditions (e.g., by assigning specific costs and values to responses). The detector responds to each new set of conditions by adjusting its threshold for calling detections. But this procedure tends to be time-consuming. Another measuring technique known as the rating method is often preferred because it simulates multiple operating conditions within a single test session, by collecting not only YES/NO decisions but also by collecting information about the level of confidence in the decisions. One can picture a link between the detection threshold and the level of confidence: with a detection threshold set to a relatively high value, the detector must be quite confident about a detection decision; lowering the threshold is equivalent to including detection calls that are made with lesser confidence. A series of measured operating points is therefore obtained by assembling and analyzing the rating-method data in different ways. The rating method was adopted for the listening tests.

The above discussion on detector performance was brief and simplistic; its purpose is to introduce the notions of the listening-test participant as a detector and of using data from a rating exercise to produce ROC curves and numeric measures of performance. Rigorous treatments of detection performance can be found in textbooks on Signal Detection Theory (SDT); the main references for this work were [5][7][8].

1.3 Report outline

The planning work for the ROC analysis is the subject of Section 2. Results of the analysis are presented in Section 3, with supporting data and results given in the annexes. A brief conclusion follows in Section 4.

2 Preparing for the analysis

2.1 Reference information

The plan for the ROC analysis influenced a number of design aspects of the listening tests, such as the number of stimuli used, the forms of data to be collected, and how they would be collected (see Section 2.2 below). Such considerations are not governed by rigid procedures or formulae; nor have they received much attention in the literature (unlike the theory behind ROC curves). The work of Swets and Pickett [7] seems to be an exception: it focuses on developing general guidelines for gathering and analyzing ROC data, and illustrates these guidelines using case studies from the field of medical diagnosis. Given that error bounds for models of observed performance usually receive inadequate attention in reports, the guidelines on the statistical treatment of accuracy were particularly useful (discussed further in Sections 2.2.4 and 2.4). The book served as the primary reference for this planning work.

The mathematical and statistical details of fitting ROC curves to experimental data were also studied during planning. This helped to evaluate the suitability of a web-based program [9] for fitting ROC curves to data (see Section 2.3 below) and to prepare for analyzing results (Section 3). Useful references found on the topic of fitting ROC curves were articles on maximum likelihood estimation applied to ROC curve-fitting [10][11] and sample software code [7][12].

2.2 Design of the listening tests

2.2.1 Rating scale

The rating-exercise task of the listening tests was used to collect ROC data; it is based on the rating method, which was introduced in Section 1.2. In this task, the listener is presented with a scale with which to rate the echo just heard. The scale format is part of the test design□ it can be continuous or discrete, labeled with numbers or descriptors. A coarse discretized scale labeled with verbal descriptors was favoured to keep the test subjects□cognitive effort to interpret the scale relatively low: the final design was the discrete seven-category (also known as seven-point) scale shown in Figure 3.[3] A comment should be made about the words *possible* and *probable* found in the descriptors: their use seemed reasonable given that the DND personnel recruited for the tests would likely be very familiar with the expressions *probable contact* and *possible contact* that are routinely used in ASW operations.

[3] Generally speaking, the number of operating points obtained from an *n*-point rating scale is *n*□*1*: therefore the greater the number of rating categories, the better the sampling of measurements across the full range of performance. Keeping in mind that the listeners should be consistent in their use of the rating categories throughout the exercise if reliable models of their discrimination ability are to be obtained, four- to seven-point scales seemed workable. The decision was made to use a seven-point scale knowing that, if required (e.g., if a listener used only a subset of the rating categories), it would be possible during the analysis to re-group the data into a smaller number of categories (see Sections 2.3.3.1 and 3.3.2).

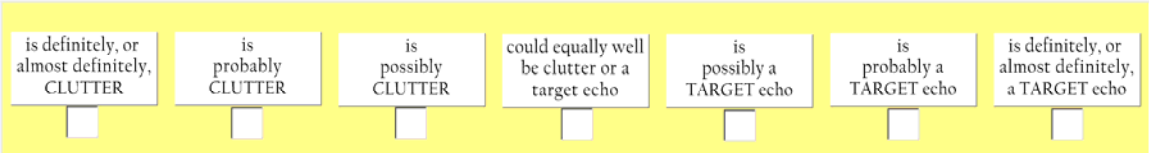

Figure 3: The seven-point rating scale used in the listening tests

As was mentioned in Section 1.2, the rating procedure associates a level of confidence to the listener's decision of target echo or clutter: the wording used for the rating scale clearly demonstrates this idea. One can also see that a decision could be made very confidently and be the wrong decision! Such hard cases for a detector (echoes that are routinely misclassified), and indeed a range of easy to hard cases for both classes of echoes, constitute measurements made at different points in the full range of performance of the detector.

2.2.2 Stimuli

The stimuli for the listening tests were obtained from recordings of an active sonar experiment involving broadband explosive sources [2][3]. With the quality of the performance models of course dependent on the quality of the data input, close attention was paid to the data-processing for generating the stimuli [3][4]. The objective was for the listeners to classify echoes based on aural properties of the echoes themselves, not on incidental cues such as position and loudness of the echoes within the sound segments. Timing and loudness of the echoes were therefore made as uniform as possible. Signal-to-noise ratio (SNR) was another consideration: the SNR of the clutter was, generally speaking, lower and more varied than that of the target echoes, which would have made it an obvious but inappropriate discrimination cue given the objective of the testing. There was, however, a way to circumvent this problem: given that the receiver for the experiment was a towed array, it was possible to make use of its inherent directionality and select off-beam target-echo signals that are of lesser strength than those received in the main beam (the one pointed directly at the target). A number of off-beam target echoes were selected so that their SNR distribution was similar to that of the clutter cases. These two sets of echoes formed the stimuli for the listening tests.

The sample size also needed to be determined. Very broad rules-of-thumb were given in Sections 4.4, 4.5, and 9.2 of [7]: a sufficient number were required for large-sample statistics to apply (a minimum of 50 was suggested), but not beyond a practical limit on how many a human subject can handle in any one test session (a number between 100 and 300 was the suggested upper limit). The final decision was to use 73 target echoes and 73 cases of clutter as the test stimuli, and another 50 echoes, 25 of each, for training. These numbers seemed manageable for a single test session, especially given that participants were given complete control over the pace of the experiment, including breaks.

2.2.3 Data collection for the original ('full-band') listening test and the follow-up ('reduced-band') listening test

There were two separate listening tests, each described by a test protocol [4], but their only significant difference lay in the bandwidth of the stimuli; the questionnaire and the rating exercises were identical for both tests. In the original test, the full bandwidth of the recorded echoes was exploited (approx. 0 □2 kHz). In the follow-up test, a high-pass filter was applied to remove any content below 500 Hz; the choice of echoes and their sequence of presentation were unchanged. The reason for the follow-up test was to investigate the effect of losing the low end of the spectrum, which, as explained in Annex B of [4], would have been a likely consequence of the sonar experiment if a realistic coherent source had been used instead of explosive sources.

2.2.4 Performance variability

There were in fact three rating exercises in the listening test; all used the rating scale discussed in Section 2.2.1. But only two of the exercises, Rating Exercise #1 and Rating Exercise #3, were designed to collect ROC data. (Rating Exercise #2, as was explained in the Addendum to the test protocols, served another purpose that is unrelated to the ROC analysis and therefore is not discussed here.) In Rating Exercise #1, each participant was asked to rate, one at a time, all 146 stimuli. Rating Exercise #3 was nearly identical: the same set of echoes were presented, but in a different sequence. The purpose of Rating Exercise #3 was to observe the variability in each participant□s performance, one of the factors taken into consideration when evaluating the reliability of a performance model. Quantifying this variability using data from the two Rating Exercises will be discussed in Section 2.4.2.

2.2.5 Qualitative data

Verbal descriptions of cues and decision processes provided by the human subjects can be useful for supporting or explaining results (Section 7.8 of [4];[13];[14]). A questionnaire and a feedback session were two techniques for collecting qualitative data that were used in the listening tests. Information about the subjects□experience with sonar and perceived aural cues were gathered in the questionnaire. The feedback session was more free-form: at the end of the tests, the experimenter would receive any comments and questions, general and specific, that the participants had about the test.

2.2.6 Preliminary results to listeners

The feedback session was designed to be two-way: to satisfy the interest and sustain the motivation of participants, the experimenter would provide them with preliminary results of their performance (Section 7.8 of [7]). The plan was to carry out an analysis of their responses to Rating Exercise #1 while they were doing the subsequent phases of the listening test and show them preliminary results of that analysis during the feedback session. The analysis was facilitated by having the program that guided the participants through the test assemble the responses into a form that was ready for analysis. This form consists of two lines of *category frequencies*: on one line, a frequency count of how many times each rating category was used to classify those echoes that were actually clutter; on the second a similar frequency count for the target echoes. An

example of the category frequencies produced by the session log is shown in Figure 4. An example of preliminary results provided to a participant during the feedback session is shown in Figure 5. An overview of the web-based program that was used to generate the preliminary results is the subject of the next section.

```
W       145     29_7_19+5_188404_fs3_NP.wav     TT      7
play sound:  9_7_79_156211_fs3_NP.wav
W       146     9_7_79_156211_fs3_NP.wav        CL      1
RS - Clutter    21      17      12      10      2       7       4
RS - Target     6       8       6       10      6       10      27
End rating test #1
Start rating test #2
Data set: May12_RE2.txt
Show cues
play sound:  31_2_60-2_155616_fs3_NP.wav
```

Figure 4: Snapshot of part of the automated session log for one of the listeners. Highlighted are the two lines of input used for the preliminary ROC analysis. A count of how many times each rating category was used to classify the cases that are actually clutter was recorded on the first line; that for the target-echo cases are given on the second line. Not shown for sake of clarity are the first three columns of information of each row (date stamp, time stamp, and identification code assigned to the listener).

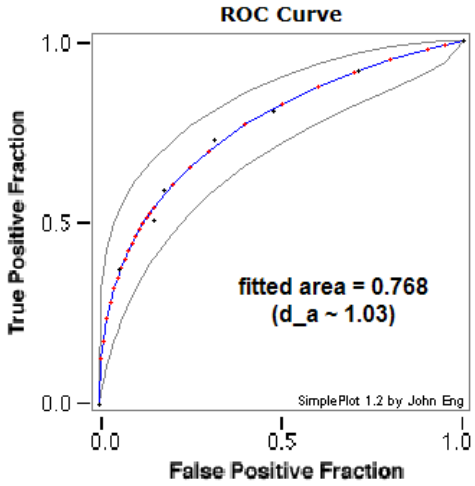

Figure 5: Sample preliminary results shown to a listener. The experimenter would go over the different parts of the graph: the set of operating points obtained from the rating answers (black dots); the fitted curve (blue curve) and sample points on the curve (red dots); the error bounds of the fitted curve (the two black curves depicting the 95% confidence interval); and the performance measure of area-under-the-curve (0.768 in this example).

2.3 Overview of a web-based program for fitting ROC curves to experimental data

A web-based program for fitting ROC curves to experimental data was available for public use [9]. It was a Java translation of an earlier Fortran program that appeared well-established and documented. It produced a ROC model for given input data, as well as error estimates on the model, and a measure of the goodness-of-fit of the model to the input data; all of which, as described in Section 15.0 of [15], are necessary for a sound statistical model. Its functionality was studied and tested as part of the preparation work for the ROC analysis of the listening-test data. An overview of the web-based program is given here, which introduces various terminology and concepts that will be used when presenting the results of the analysis.

2.3.1 Model to be fitted

The starting point is the assumption that the target-echo sounds and the clutter sounds as perceived by a listener can be modeled as two normal probability density functions (pdf's). The relative properties of the pdf's, i.e. the distance between their means and the ratio of the standard deviations (or variances) define the listener's ability to discriminate between the two classes of events (this is the notion of detector sensitivity that was introduced in Section 1.2). It is usual in signal detection theory to scale the pdf's such that the mean and standard deviation of the clutter pdf are 0 and 1, respectively. Modeling detector sensitivity therefore amounts to estimating the mean and standard deviation of the target pdf (see Figure 6). Also usual is the labelling of the x-axis relative to the clutter pdf, i.e., in terms of the number of standard deviations to the right and left of the 0 mean of the clutter pdf. Before looking at the second set of unknowns of the model in Figure 6 (the position of the thresholds on the x-axis, which are related to the operating points of a detector, a notion also introduced in Section 1.2), the link between the two pdf's and the corresponding swept binormal ROC curve (the qualifier "binormal" used to emphasize the underlying assumption) is briefly reviewed using the example in Figure 7.

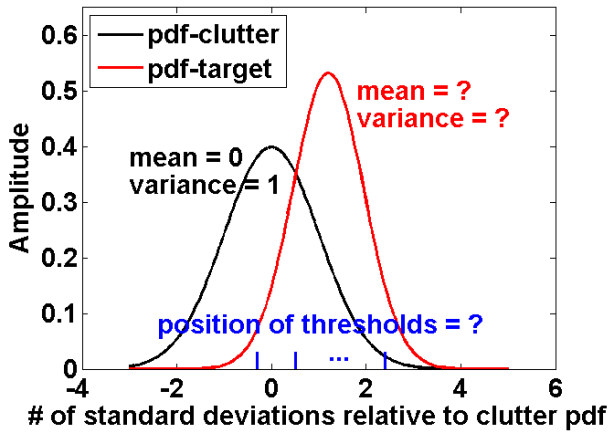

Figure 6: The assumed model with the unknown parameters: the mean and standard deviation of the target pdf and the positions of the decision thresholds.

The two pdf's shown on the left side of Figure 7 have known parameters. By definition, the probability of detection (PD) and probability of false alarm (PFA) for a given value of x, the term threshold is used here to represent of setting of a minimum value of x above which all events will be classified as target events, corresponds to the areas under the target pdf and clutter pdf, respectively, to the right of this threshold [5]. These values of PD and PFA correspond to one point on a binormal ROC curve. A full binormal ROC curve is swept by smoothly varying the position of the threshold across the breadth of the pdf's, to asymptotically reach the (0,0) and (1,1) extremities of the curve.

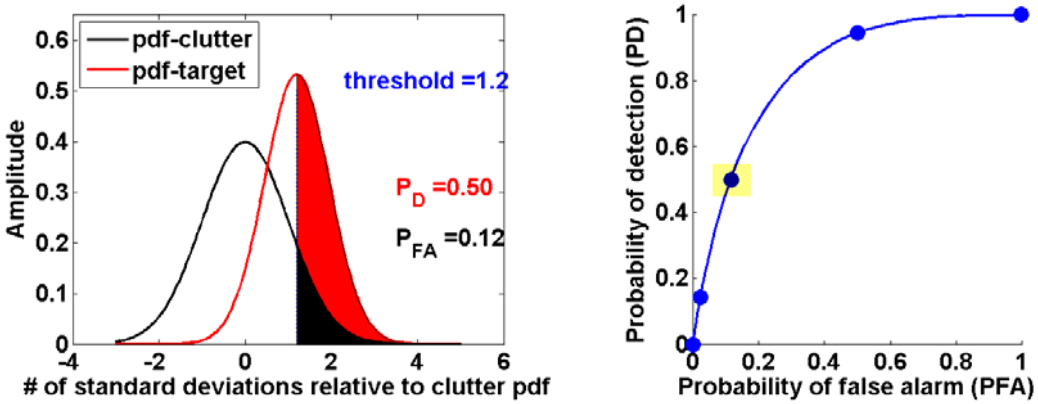

Figure 7: Illustrating the relationship between the two normal pdf's and a swept binormal ROC curve. (In this example, the mean μ and standard deviation σ of the target pdf are, respectively, 1.2 and 0.75; those for the clutter pdf are, respectively, 0 and 1.) The coordinates (0.12, 0.5) of the highlighted point on the ROC curve are obtained by fixing the threshold (in this example the threshold is set to 1.2) and calculating the area under each pdf to the right of the threshold, i.e., summing all the detection events (correct and incorrect). The full ROC curve can be obtained by smoothly varying the threshold value.

In the assumed model presented in Figure 6, the unknowns include the positions of a finite number of thresholds. (Following the above illustration of the link between the pdf's and the binormal ROC curve, determining the position of these thresholds is equivalent to determining specific points on the ROC curve known as operating points, their mathematical relationship is described in Section 2.3.3). These values arise because of the finite rating scale used in the listening tests. The number of rating categories used by the listener determines the number of operating points. If one disregards the extreme threshold positions (set extremely high such that no detection calls are made, the (0,0) point on the ROC curve, or set extremely low such that all events are called detections, the (1,1) point on the ROC curve), six *measured* threshold positions (or, equivalently, six measured operating points) are obtained when all seven rating categories are used. The number decreases by one for every rating category not used. When fitting a binormal ROC curve to these measurements, a set of modeled thresholds (set of operating points lying on the curve) will also be produced. Information on the numerical procedure used by the program to fit the model to rating-scale data is provided in Section 2.3.3.

The underlying assumption of binormal pdf's is typical in ROC analysis; but it is noted that other distributions have also been used (Chapter 3 of [5]). The Gaussian assumption has theoretical support from the central limit theorem. But perhaps even more important is its mathematical convenience: a transformation of coordinates converts the binormal ROC curve into a straight line. The fitting of the model therefore becomes a matter of fitting a straight line to data (one of the steps of the numerical procedure to be described shortly). Another convenience of the Gaussian assumption is that the error estimates can be stated and understood in terms of the usual statistical notions of variance, standard deviation, and confidence limits.

2.3.2 Program input

The program accepts input data in a number of formats; the arrangement of the two lines of data recorded in the session log (see Figure 4) corresponds to Format 3. The number of rating categories is also required.

2.3.3 Fitting procedure

The main steps of the fitting procedure are summarized in Sections 2.3.3.1 to 2.3.3.6.

2.3.3.1 Converting the category frequencies to observed operating points

The rating-scale data are transformed into a set of observed operating points. This is achieved by successively considering each boundary between rating categories as a decision threshold and summing the detection events to the right of the threshold. The calculation for one of the observed operating points is illustrated in Table 1.

Table 1: Showing one of the positions for the detection threshold and the corresponding observed operating point on a ROC graph. In this position, all events that were rated "possibly a target echo", "probably a target echo", or "definitely a target echo" are considered to be detections. The incorrect detections and correct detections are summed separately to provide the probability of false alarm (PFA) and probability of detection (PD) coordinates of the observed operating point. The numbers used for this example are based on the rating-scale data shown in Figure 4.

	definitely clutter	*probably clutter*	*possibly clutter*	*don't know*	*possibly a target echo*	*probably a target echo*	*definitely a target echo*
proportion of clutter	0.29	0.23	0.16	0.14	**0.03 +**	**0.10 +**	**0.05**
proportion of target echoes	0.08	0.11	0.08	0.14	**0.08 +**	**0.14 +**	**0.37**
(PFA,PD)					**(0.18,0.59)**		

2.3.3.2 Transforming the binormal ROC curve into a straight line

The observed operating points are converted from probabilities to standard normal deviates, also known as z-scores. The relationship between a probability and its corresponding z-score is given by

$$P(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt, \tag{3}$$

where the z-score is the upper limit of the integral. A binormal ROC curve becomes a straight line when it is plotted in terms of z-scores rather than probabilities[4] (rational approximations exist for calculating the z-score given a probability value□ see Chapter 26 of [6]). Thus the problem now becomes one of fitting a straight line to data. The slope *s* and ordinate *b* of the fitted line are related to the pdf parameters as follows:

$$s = \sigma_{\text{clutter}} / \sigma_{\text{target}} \tag{4}$$

$$b = | \mu_{\text{target}} - \mu_{\text{clutter}} | / \sigma_{\text{target}} \tag{5}$$

Figure 8 shows the z-scores of the PFA and PD components of the observed operating points for the example used in Figure 4 and Table 1.

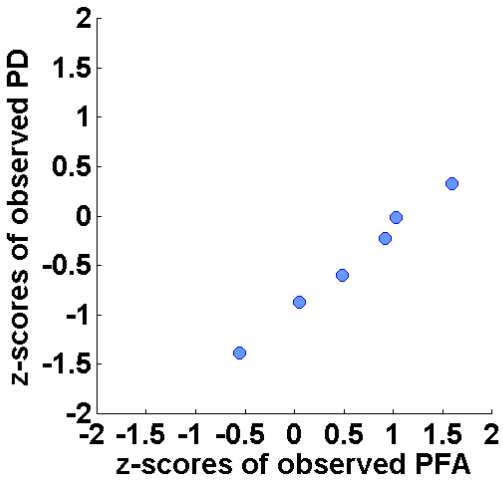

Figure 8: Using z-scores to represent the observed operating points. The sample points shown were obtained from the input data shown in Figure 4 and Table 1.

[4] Strictly speaking, the probabilities are Q(z), not P(z), where Q(z) = 1 □P(z). This extra step would lead to sign changes in the z-scores and the ordinate of the line. But the web-based program treats the probabilities as P(z) values, does the line-fitting, and handles the sign switch when producing the output information.

2.3.3.3 Initial estimate of the model parameters

The method of least-squares is applied to obtain an initial estimate of the slope and ordinate of the fitted line. The observed PFA values are used for the initial estimate of the fitted operating points. This initial fit is illustrated in Figure 9 for the sample data in Figure 8.

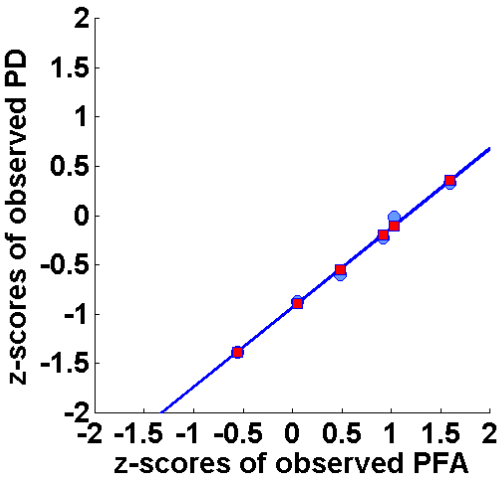

Figure 9: Initial least-squares fit to the points shown in the previous figure. The red squares represent the initial estimate of the fitted operating points.

2.3.3.4 Iterative procedure to produce maximum-likelihood estimates of the model parameters

The parameter estimates are subsequently refined using an iterative procedure that seeks to maximize a figure-of-merit function. The merit function used by the program is the log-likelihood which, for rating-scale data, takes on the form

$$\log L = \sum_{i=1}^{2} \sum_{j=1}^{n} r_{ij} \log\left(P_{ij}\right) \quad , \tag{6}$$

where n is the number of rating categories, r_{ij} is the observed number of responses in the j^{th} rating category for the i^{th} stimulus class (two classes clutter and target echo), and P_{ij} is the computed probability (based on model parameters) of using the j^{th} rating category for an i^{th} class of stimulus. Maximizing the log-likelihood leads to maximum-likelihood estimates of the model parameters.

Successive estimates of the parameters are obtained by applying Fisher s method of scoring, which is a variation on the Newton-Raphson method. The governing equation is of the form

$$\mathbf{S}^{(\mathbf{n}+1)} = \mathbf{S}^{(\mathbf{n})} + M^{-1}\mathbf{r} \quad , \tag{7}$$

where $\mathbf{S}^{(n+1)}$ and $\mathbf{S}^{(n)}$ are the vectors of parameters θ at the n+1th and nth iterations, respectively, $\mathbf{r}$ is the vector $\partial(\log L)/\partial\theta$, and M is the Fisher information matrix, for which the elements are defined as

$$M_{ij} = -E\left(\frac{\partial^2(\log L)}{\partial\theta_i\partial\theta_j}\right) \quad . \tag{8}$$

The vector $\mathbf{r}$ and matrix M are evaluated using parameter estimates at the nth iteration. The parameters consist of the z-scores of the PFA components, and the slope and ordinate of the fitted line; the number of parameters is therefore the number of category frequencies that are used plus one. The expectation – the $E(\ )$ in Equation (8) – is taken with respect to the category frequencies, which eliminates some of the terms that would ordinarily have been taken into account if the Newton-Raphson method had been applied. Furthermore, the final M^{-1} represents the large-sample variance-covariance matrix of the maximum-likelihood estimates of the parameters, which conveniently provides the numbers necessary for producing an estimate of the likely errors on the parameters. These error estimates are discussed further in Section 2.3.3.5.

The iterative procedure provides a solution vector in z-score space that consists of the ordinate and slope of the ROC line and the z-scores of the probability of false alarm of the operating points. Equations (3)-(5) can be used to obtain more usual parameters for describing ROC performance, such as the parameters of the pdf's, points on the ROC curve, and derived numeric indexes of performance. One such index is the area under the fitted binormal ROC curve, which was first introduced in general terms in Section 1.2. The symbol A_z will be used in this report to designate this index. Because of the binormal assumption, this index can be expressed directly as a z-score, using

$$z(A_z) = \frac{b}{\sqrt{1+s^2}} \tag{9}$$

or stated as a proportion of the total ROC space that lies under the curve, either by using Equations (3) and (9) or by actually calculating the area under the curve.

2.3.3.5 Error estimates

The iterative technique outlined in Section 2.3.3.4 produces best-fit parameters and, in the process, generates the large-sample statistics of the parameters in the form of the variance-covariance matrix. The fitting procedure therefore provides a measure of the accuracy with which parameters are determined by the experimental data, i.e. provides error estimates on the parameters.

Confidence intervals are typically used when graphically depicting the amount of uncertainty associated with statistical estimates. The program uses the matrix information to calculate confidence intervals for the fitted ROC curve and the fitted operating points on the curve.

When calculating the asymmetric 95% confidence intervals for the ROC curve, it is assumed that errors in the ordinate and slope estimates are jointly normal. The upper and lower bounds of the confidence intervals are worked out in z-score space, then converted to probabilities using Equation (3). Using the symbols z_{PFA} and z_{PD} to denote the z-score of a probability of false alarm and the z-score of a probability of detection, the upper and lower bounds of a confidence interval are calculated for a given set of values of z_{PFA} using

$$z_{PD} \pm 1.96\sigma(z_{PD}) \quad , \tag{10}$$

where

$$\sigma(z_{PD}) = \sqrt{\mathrm{var}(b) + \mathrm{var}(s) \times z_{PFA}^2 + 2 \times z_{PFA} \times \mathrm{cov}(s,b)} \ . \tag{11}$$

When calculating the asymmetric 95% confidence intervals for the fitted operating points, it is assumed that the errors in the z-scores of the probability of false alarm-component of the fitted operating points, z_{PFA_OpPt}, are normally distributed. The bounds are calculated using

$$z_{PFA\_OpPt} \pm 1.96\sigma(z_{PFA\_OpPt}) \tag{12}$$

and

$$z_{PD\_OpPt} = s \times z_{PFA\_OpPt} + b \quad . \tag{13}$$

An expression for the standard deviation of the z-score of the index A_z, $\sigma(z(A_z))$ can be obtained by using Equation (9) and the relationship

$$\mathrm{var}(z(A_z)) = \left(\frac{\partial z(A_z)}{\partial b}\right)^2 \mathrm{var}(b) + \left(\frac{\partial z(A_z)}{\partial s}\right)^2 \mathrm{var}(s) + 2\frac{\partial z(A_z)}{\partial b}\frac{\partial z(A_z)}{\partial s}\mathrm{cov}(s,b) \tag{14}$$

The result is easily transformed into σ_{Az} by multiplying by y^2, where y is the value of the probability density corresponding to the z-score. The confidence interval of the index A_z may be reasonably estimated using

$$A_z \pm 1.96\sigma_{Az} \tag{15}$$

provided the estimate of A_z is not too close to the upper limit of unity; Section 4.5 of [7] briefly discusses the problem of the errors not being normally distributed for these extreme cases and possible measures to address the problem.

2.3.3.6 Measure of goodness-of-fit

The program's fitting procedure includes a goodness-of-fit test for measuring how well the model predictions agree with the experimental data. Perfect agreement is not expected except in the case where the listener uses only three of the seven rating categories – in this situation, the problem becomes one of fitting a straight line to two points, which can be done exactly.

The statistical goodness-of-fit test that is implemented is the standard Pearson's chi-square test (descriptions of the test are readily found; see for example Section 14.3 of [15] or Section 13.1 of [16]). The test applies to binned data. The input data, the observed category frequencies, are already in this form; expected category frequencies can be calculated from the fitted operating points of the model. The test statistic is defined as

$$X^2 = \sum_{i=1}^{k} (O_i - Np_i)^2 / Np_i \tag{16}$$

where O_i is the observed number of events in the ith bin and Np_i is the expected number (N is the total number of events and p_i is the expected probability associated with the ith bin). The web-based program calculates this statistic, the number of degrees of freedom (the number of effective operating points minus two given that the slope and ordinate parameters are calculated from the points), and the probability of encountering this statistic assuming that it follows a chi-square distribution. If the probability is deemed significant (smaller than 5%, say), the model can be rejected as unsuitable for representing the experimental data.

It should be noted that the chi-square test does not always produce reliable results. The test is not recommended when the sample size is small or when the expected binned frequencies are small. An often quoted rule-of-thumb for judging if the chi-square test is suitable is that all expected binned (category) frequencies must be at least five (Section 13.1 of [16]). This rule-of-thumb is also used by the program: it only runs the chi-square test when this condition is met, and produces a warning message when it is not met.

2.3.4 Program output

The program produces result summaries in graphical form and in text form. The contents of two of the output windows, *ROC Curve* and *Program Output,* were very relevant to the listening test and subsequent analysis. The graphical display of the fitted ROC curve, associated asymmetric 95% confidence intervals, and observed operating points was presented as part of the feedback to test participants (see Section 2.2.5). The contents of the *Program Output* window were copied and saved as a text file for use in the analysis. This output contains a record of

- the input data;
- the observed operating points;
- initial and final estimates of the log-likelihood function and the parameters;
- the final variance-covariance and correlation matrices;

- the resulting A_z index of performance and estimated standard deviation of A_z;
- coordinates of points representing the fitted ROC curve and associated asymmetric 95% confidence intervals;
- coordinates of points representing the expected operating points and associated asymmetric 95% confidence intervals; and
- if carried out, results of the χ^2 goodness-of-fit test.

2.3.5 Accuracy of the program output

The program was used for a number of tests to learn about its functionality and to check its reliability (the owners of the program had made it readily available but did not guarantee accurate operation). All but one output value, the probability calculation for the chi-square measure of goodness-of-fit, seemed consistently reasonable. Easy workarounds for obtaining an accurate probability are to use a look-up table or to write a computer program (a chi-square test was implemented in MATLAB® for this ROC analysis).

2.4 Measures of performance

A number of measures have been used to describe detector performance [5][7]. The one chosen for this analysis (and recommended in [7]) is based on the numeric index A_z, which corresponds to the area under the binormal ROC curve. As described in Section 2.3, when the program fits a binormal ROC curve to input data, it provides, as part of its output, estimates of A_z and the standard deviation of A_z, σ_{Az}. These output values for individual rating exercises can be assembled to provide measures of performance of individual listeners and groups of listeners, and are easily depicted using confidence intervals. Furthermore, a statistical test can be applied to the measures of performance to determine if the differences are significant.

The components of variance that are considered when determining the standard error of the measure of performance are discussed in general terms in Section 2.4.1. A procedure for estimating the average performance of a group of listeners is summarized in Section 2.4.2, followed by a proposed procedure for estimating the performance of individual listeners in Section 2.4.3. The intended statistical test for comparing measures of performance is presented in Section 2.4.4.

2.4.1 Components of variance

Performance variability is discussed here in general terms: a more thorough account can be found in Chapters 3 and 4 of [7]. The value of A_z reported by the web-based program is a maximum-likelihood estimate of performance of a particular listener for a particular sample of target echoes and clutter for one particular occasion (one of the Rating Exercises). The reported value of σ_{Az} is a maximum-likelihood estimate of the amount of variability in A_z that would be observed if the listener had carried out the Rating Exercise on different samples of stimuli, each sample taken

from the same population of target echoes and clutter from which the experimental data set of echoes was taken. Embedded in this observed variability is the listener's inconsistency in classifying echoes (the listener will not always rate a given echo the same way on every occasion). The value of σ_{Az} is therefore a combined measure of the variability from sample to sample and the variability in the listener's own performance. This idea can be expressed mathematically as

$$\sigma_{Az} \approx \sigma_{s+wl} = \sqrt{\sigma_s^2 + \sigma_{wl}^2} \quad , \tag{17}$$

where σ_s^2 is the sample variance and σ_{wl}^2 is the within-listener variance. In asking the human subject to repeat the Rating Exercise, it becomes possible to estimate the two components of variance individually. This is further discussed in Section 2.4.2.

Performance between the listeners will also vary. Differences in the reported values of A_z for a group of listeners will be due to the varying ability among the listeners to distinguish target echoes from clutter and the variability in each listener's own performance. This observed variability is expressed mathematically as

$$\sigma_{bl+wl} = \sqrt{\sigma_{bl}^2 + \sigma_{wl}^2} \quad , \tag{18}$$

where σ_{bl}^2 is the between-listener variance. Given that σ_{wl}^2 has already been estimated, Equation (18) can be used to estimate σ_{bl}^2.

2.4.2 Measure of average performance of the human listeners

A procedure for producing a measure of average performance of a group of test subjects was presented in Section 4.1 of [7]. It was adopted for this analysis. A summary is given below.

Average performance achieved by a group of listeners is statistically described by a mean value A_{z_group} and standard error of the mean (σ_{Az_group}). The standard error takes into account the sample, between-listener, and within-listener variances introduced in Section 2.4.1. The basic equation linking these components of variance to the standard error is given by

$$\sigma_{Az\_group} = \sqrt{\sigma_s^2 + \frac{\sigma_{bl}^2}{l} + \frac{\sigma_{wl}^2}{lm}} \quad , \tag{19}$$

where l is the number of listeners and m is the number of times the Rating Exercise is carried out by each listener. A case is made in Section 4.1 in [7] to simplify the calculations and to use an approximation to Equation (19) which states the error in terms of σ_{s+wl}^2, σ_{bl+wl}^2, and σ_{wl}^2. The equation proposed is

$$\sigma_{Az\_group} = \sqrt{\sigma^2_{s+wl} + \frac{\sigma^2_{bl+wl}}{l} - \sigma^2_{wl}} \quad . \tag{20}$$

The three variance components in Equation (20) are estimated as follows:

- the arithmetic mean of the l values of σ^2_{Az} obtained for Rating Exercise #1 is used as an estimate of σ^2_{s+wl};
- the biased variance (second moment) of the l values of A_z obtained for Rating Exercise #1 is used as an estimate of σ^2_{bl+wl}; and
- for each listener, a biased variance (second moment) of the values of A_z associated with that listener is calculated. The arithmetic mean of all these variances is used as an estimate of σ^2_{wl}.

The arithmetic mean of the l values of A_z obtained from Rating Exercise #1 is used as an estimate of A_{z_group}. It is an estimate of the average performance of DND personnel with significant experience in sonar in classifying the target echoes and clutter represented by the sample of echoes used in the listening tests.

2.4.3 Measure of average performance of an individual listener

The idea here is to produce a measure of performance for each human listener and compare it to that of the automatic classifier with a view to gauging the rank of the "machine" participant's performance in the listening tests. A measure of average performance of listener X is proposed. It consists of a mean value, A_{z_X}, and standard error of this mean, σ_{Az_X}, where the arithmetic mean of the *m* values of A_z associated with listener X is used as an estimate of A_{z_X}, and the standard error is estimated using

$$\sigma_{Az\_X} = \sqrt{(\sigma^2_{s+wl\_1} + \sigma^2_{s+wl\_2})/2} \quad . \tag{21}$$

2.4.4 Comparing measures of performance

A z-test (Section 7.4 of [17]) was suggested in [7] as a method for determining if two measures of average performance differ significantly. The z-statistic is approximated as

$$z_{stat} = \frac{A_{z\_1} - A_{z\_2}}{\sigma_{Az\_1-Az\_2}} \quad , \tag{22}$$

where the numerator is the difference in the means and the denominator is the standard error of the difference. One then refers to a standard normal curve to find the probability that $z \geq | z_{stat} |$. If the significance level is 5% say, then the difference between the two measures of performance is deemed statistically significant. Alternatively, one can state the significance level as a critical value of z (for 5%, $z_{cv} = 1.96$) and call the result significant when $| z_{stat} | \geq z_{cv}$. Equation (22) is an approximation because it uses the standard deviation associated with the classification performance of a single sample of echoes rather than the standard deviation associated with the entire population of echoes from which the sample was taken. The amount of work associated with repeating the tests to sample the entire population is usually unfeasible. But the approximation was deemed reasonable in the case where the sample size is considered to be large (Section 4.5 of [7]). This constraint was considered when designing the test (see Section 2.2.2) so that the approximation could reasonably be applied in this analysis.

The numerator in Equation (22) is straightforward. When the two means vary independently, the denominator is expressed as [18]

$$\sigma_{Az\_1-Az\_2} = \sqrt{\sigma^2_{Az\_1} + \sigma^2_{Az\_2}} \quad . \tag{23}$$

But the standard error of the difference will likely be smaller because the two means do not vary independently: there is some correlation in the variance components that are used to estimate the standard errors of the means (Sections 3.3 and 4.3 of [7]). A smaller standard error would increase the power of the statistical test [7]. Given that the interest here is in comparing the performance of the human listeners and the automatic classifier, the only correlation possible is in the sample variance: the between-listener and within-listener variances for the automatic classifier are zero since, for a given training set of echoes, it consistently produces the same results for the testing echoes (those used for the Rating Exercises). The method proposed in Section 4.3 of [7] to estimate the correlation in the sample variance was not judged suitable for this analysis because the sample of echoes used in the listening tests is not sufficiently large. Rather than try to guess a suitable number for the correlation or develop another method to estimate it, the plan for comparing performances was to use the conservative test based on Equation (23).

3 Analysis

3.1 ROC input data from the listening tests

There were fifteen human subjects for the full-band test and ten for the reduced-band test. Nine participated in both tests. The responses to Rating Exercise #1 and Rating Exercise #3 for all the subjects except one were considered for the ROC analysis. A different set of training and testing echoes were inadvertently used for test participant *s02*: given the good turnout of volunteers for the tests, it seemed reasonable to discard *s02*'s responses to the stray data set, and base the analysis on the data from the remaining fourteen and nine subjects (with eight participating in both tests), who were all trained and tested on the same sets of echoes. All volunteers were military and civilian DND personnel with significant experience in sonar.

The input data for the ROC analysis consist of the *category frequencies* of the subjects' responses. The term was first used in Section 2.2.6 to refer to a frequency count of how many times each rating category was used to classify those echoes that were actually clutter and a similar frequency count for the echoes that were actually target echoes. The category frequencies are found in Annex A.

3.2 A rough first analysis of the human-performance data

The web-based program that was described in Section 2.3 was used to generate performance results of each human subject in Rating Exercise #1 and Rating Exercise #3. Each result contains three values: the index of performance area A_z (representing the area under the fitted binormal ROC curve), the standard deviation of A_z, and a measure of goodness-of-fit of the model to the input data. Results for the full-band test are presented in Table 2; those for the reduced-band test in Table 3. Also included are measures of average human performance achieved by the listeners (calculated using the method described in Section 2.4). Comments on the results are found below.

Table 2: Results of the first analysis of the full-band data. The result describing an individual's performance in one of the rating exercises consists of a value A_z, a value of the standard deviation of A_z (σ_{Az}), and a measure of goodness-of-fit (Q). Also shown is a measure of average human performance achieved for the full-band test.

Listener code	Rating Exercise #1			Rating Exercise #3		
	A_z	σ_{Az}	Q	A_z	σ_{Az}	Q
s01	0.99	0.01	n/c*	--□	--	--
s03	0.91	0.02	n/c	0.97	0.01	n/c
s04	0.99	0.01	n/c	0.99	0.01	n/c
s05	0.97	0.01	n/c	0.96	0.01	n/c
s06	--	--	--	--	--	--
s07	0.77	0.04	n/c	0.82	0.04	n/c
s08	1.00	0.00	n/c	0.99	0.01	n/c
s09	0.95	0.02	n/c	0.96	0.01	n/c
s10	0.98	0.01	n/c	0.98	0.01	n/c
s11	0.83	0.04	n/c	0.88	0.04	n/c
s12	0.98	0.01	n/c	0.97	0.01	n/c
s13	0.96	0.02	n/c	0.99	0.01	n/c
s14	0.97	0.01	n/c	0.94	0.02	n/c
s15	0.95	0.02	n/c	0.98	0.01	n/c
average performance	$A_{z\,mean}$ = **0.94**			$\sigma_{Az\,mean}$ = **0.02**		

* Not computed. The chi-square test of goodness-of-fit was not executed by the web-based program because one or more expected category frequencies were less than five (see Section 2.3.3.6).

□ No values were produced because no realistic binormal ROC curve could be fit to the data. This occurs when the subject uses no more than two rating categories for rating the target-echo cases or the clutter cases.

Table 3: Results of the first analysis of the reduced-band data. The result describing an individual's performance in one of the rating exercises consists of a value A_z, a value of the standard deviation of A_z (σ_{Az}), and a measure of goodness-of-fit (Q). Also shown is a measure of average human performance achieved for the reduced-band test.

Listener code	Rating Exercise #1			Rating Exercise #3		
	A_z	σ_{Az}	Q	A_z	σ_{Az}	Q
s01	0.65	0.05	n/c*	0.67	0.05	n/c
s04	0.70	0.04	n/c	0.71	0.04	n/c
s06	0.79	0.05	1	--	--	--
s08	0.91	0.03	n/c	0.90	0.03	n/c
s09	0.76	0.04	n/c	0.68	0.05	n/c
s10	0.69	0.04	n/c	0.75	0.04	n/c
s12	0.82	0.04	n/c	0.76	0.04	n/c
s15	0.75	0.04	n/c	0.73	0.04	n/c
s16	0.87	0.03	n/c	0.85	0.04	n/c
average performance	$A_{z\,mean}$ = **0.77**			$\sigma_{Az\,mean}$ = **0.04**		

* Not computed. The chi-square test of goodness-of-fit was not executed by the web-based program because one or more expected category frequencies were less than five (see Section 2.3.3.6).

 Perfect fit. The input data provided only two operating points, to which a binormal ROC curve could be fit exactly. A goodness-of-fit test was therefore not needed.

No values were produced because no realistic binormal ROC curve could be fit to the data. This occurs when the subject uses no more than two rating categories for rating the target-echo cases or the clutter cases.

Remarking first on the general trend of the values of A_z, one can see that they all substantially exceed the 0.5 index of chance performance (introduced in Section 1.2), indicating that the human subjects were hearing differences in the echoes that allowed them to make reasonably good classification decisions. Also, the values for the full-band test are higher than those for the reduced-band test, indicating that when the band 0 500 Hz was removed, aural differences in the echoes became less obvious. These observations are made without considering the reported values of standard deviation, which statistical testing would also take into account when comparing values of A_z. But before proceeding to such tests, there are two irregularities in the results that should be addressed.

The first is that four sets of results are missing from the above tables: participant *s01*'s performance in Rating Exercise #3 of the full-band test, participant *s06*'s performance in Rating Exercises #1 and #3 in the full-band test, and participant *s06*'s performance in Rating Exercise #3 in the reduced-band test. Results were not produced because no realistic binormal ROC curves could be fitted to the input data. This happened because the test subject used no more than two rating categories when classifying the cases that were target echoes or the cases that were clutter. Possible explanations for this trend were that *s01*'s and *s06*'s hearing ability is different from the others or that they are using different aural cues to make their decisions. There was, however, no obvious evidence for either of these explanations when looking over the qualitative data collected in the tests. Both listeners, however, were very open about their preference for a black-and-white decision-making strategy: either it is a target echo or it is not. (It can be seen from the data recorded in Annex A that these participants favoured using the extreme categories of the rating scale.) It should be said that the experimenter did not tell subjects how to use the rating scale; rather, the point was made to them that they should use the scale as they saw fit. Their decision-making strategy was not invalid or wrong: it simply did not provide the data that were necessary to obtain a realistic value of A_z. Given this observation, the approach used to deal with the missing results when calculating measures of average performance does not seem unreasonable.

The second irregularity is that not a single measure of goodness-of-fit was produced. (And only in one case was the goodness-of-fit test not necessary because the human subject used only three rating categories, which produced two operating points, to which a binormal ROC curve could be fitted exactly hence the '1' entry in Table 3) As mentioned in Section 2.3.3.6, the standard chi-square test was implemented in the web-based program to measure the goodness-of-fit of the binormal ROC curve to the input data. This test, however, is recognized as being unsuitable when one or more category frequencies are small. Rather than produce unreliable results, the web-based program does not run the test and produces a warning message for the user. Unfortunately, this was the case for all the models produced.

Without a measure of how well the models represent the measured data, the measures of performance can be questioned. Ways to assess the goodness-of-fit are considered in a more detailed analysis of the human-performance data, which is the topic of Section 3.3.

3.3 Further analysis of the human-performance data

Two approaches for assessing the goodness-of-fit of the models to the input data were explored. The first approach was to adopt a goodness-of-fit test that would produce reliable results when category frequencies are small. The second was to combine categories frequencies so that it would be reasonable to apply the chi-square test. Results of each approach, including effects on the measures of human performance, are discussed separately in Sections 3.3.2 and 3.3.1.

3.3.1 Randomization test of goodness-of-fit

Unlike the chi-square test, a randomization test [19] of goodness-of-fit does not assume a distribution for the test statistic of interest, so its reliability is not compromised when the category frequencies are small. The idea behind the randomization test is to generate the test statistics for a very large number of samples of modelled data, i.e. produce an actual distribution of the test

statistic, and see where the observed statistic (the one that represents the measured data) fits in. If the probability of encountering the observed statistic in the distribution is very small (5% is a typical threshold used in statistics), the model is deemed ill-suited to represent the measured data.

3.3.1.1 Implementation

A randomization test of the test statistic was implemented in MATLAB®. The program steps are the following:

1. Calculate the expected category frequencies from the fitted operating points of the model.
2. Using Equation (16), calculate the test statistic associated with the measured data.
3. Generate two vectors, one representing the modelled target-echo population, the other representing the modelled clutter population. Each vector will contain the numbers 1 to 7, each number designating a rating category, in proportions that are consistent with the expected category frequencies. For example, if the length of the vector for the clutter case is N, the number i (with $1 \le i \ge 7$) appears in the vector n_i times, where $n_i = Np_i$ and p_i is the expected probability associated with the ith category of the modelled clutter population.
4. Use the bootstrap method (Section 15.6 of [15]) to generate an actual distribution of the test statistic and calculate the probability of encountering the observed statistic in this distribution. This step is further detailed as follows:

    a. Draw a random target-echo sample and a random clutter sample from the population vectors. For example, a random clutter sample is produced by selecting at random, one at a time, N elements of the clutter population vector and counting the resulting number of 1's, 2's, ..., 7's obtained, i.e. counting the category frequencies in this random sample.

    b. Calculate and add the test statistics for the generated target-echo and clutter samples. Identify if the result is larger or smaller than the statistic associated with the measured data.

    c. Repeat steps a and b a large number of times – in this implementation, 10000 samples of target echoes and 10000 samples of clutter were generated. Count the number of samples that produced a simulated test statistic that was equal or larger than the statistic associated with the measured data. State as a proportion of the total number of samples. This is the measure of goodness-of-fit *(Q)*.

5. If *Q* is smaller than the set threshold (5% will be used), the result is deemed statistically significant, that is to say the model is ill-suited for representing the measured data. (Using statistics jargon, a statistically significant result means that the null hypothesis can be rejected, where the null hypothesis is that the measured data are consistent with the model.)

3.3.1.2 Goodness-of-fit results

Results of applying the randomization test of goodness-of-fit to the full-band models are presented in Table 4; those for reduced-band models are found in Table 5.

Table 4: Results of applying the randomization test of goodness-of-fit to the full-band models. Shown are the observed test statistics (X^2) and measures of goodness-of-fit (Q). Two of the results, shown in bold, are deemed statistically significant.

Listener code	Rating Exercise #1		Rating Exercise #3	
	X^2	Q	X^2	Q
s01	0.6	1.00	--[1]	--
s03	78.2	**0.00**	4.1	0.97
s04	6.5	0.89	4.4	0.99
s05	3.5	0.99	8.2	0.78
s06	--	--	--	--
s07	1.8	1.00	12.2	0.43
s08	3.2	0.99	4.6	0.95
s09	21.1	**0.04**	12.4	0.41
s10	3.5	0.94	0.4	1.00
s11	2.6	0.99	1.9	0.96
s12	4.7	0.94	10.0	0.62
s13	7.6	0.65	2.6	0.99
s14	9.3	0.59	6.3	0.79
s15	0.7	1.00	2.0	1.00

[1] No model fitted to data – see footnote in Table 2.

Table 5: Results of applying the randomization test of goodness-of-fit to the reduced-band models. Shown are the observed test statistics (X^2) and measures of goodness-of-fit (Q). None of the results are deemed statistically significant.

Listener code	Rating Exercise #1		Rating Exercise #3	
	X^2	Q	X^2	Q
s01	2.6	0.96	2.1	0.98
s04	3.4	0.99	6.6	0.88
s06	0[1]		--[2]	--
s08	2.5	1.00	1.3	1.00
s09	3.4	0.97	3.8	0.87
s10	6.9	0.73	4.1	0.95
s12	4.3	0.99	2.6	1.00
s15	10.8	0.54	4.9	0.96
s16	3.1	1.00	11.5	0.49

[1] Perfect fit. The input data provided only two operating points, to which a binormal ROC curve could be fit exactly. A goodness-of-fit test was therefore not needed.

[2] No model fitted to data – see footnote in Table 2.

3.3.1.3 Possible reasons for the statistically significant results

There are two statistically significant results: both are associated with full-band models. Qualitative data collected during the listening tests were examined for evidence that the misfit between the model and the measured data is due not only to random fluctuations.

The first significant result is for subject *s03*'s performance in Rating Exercise #1. A possible reason for the peculiar performance may be inferred from the subject's answers to the questionnaire. This listener had several years of experience in passive sonar and was very used to making Doppler calls – and seemed intent on applying this experience to the listening test, which would be very misleading given that the stimuli were active sonar echoes. However, the participant did note during training that target echoes had clarity and sharpness qualities that the clutter did not have. Given that a significant value of Q was obtained for Rating Exercise #1 but

not for Rating Exercise #3, it may be that the listener used some perceived Doppler sound as the main cue for Rating Exercise #1 but relied on it less for Rating Exercise #3.[5]

The second significant result is for subject *s09*'s performance in Rating Exercise #1. Judging by the answers to the questionnaire and the comments during the feedback session, it seems that the listener was very keen on applying cue(s) that had been learned and used during the several years of experience with both active and passive sonar, preferring these to potential cues that had been identified during the training portion of the listening test (target echoes being more intense and crisp than the clutter; a double echo or reverb sound being a sign of clutter). A conflict in cues or hesitancy to use cues may explain why this listener reported not being confident when classifying some of the echoes. A switch in the importance of cues could explain why only the result for Rating Exercise #1 was significant.

Subject *s09* did return to participate in the reduced-band test, and qualitative data collected seems to suggest that the listener was less hesitant to apply cue(s) perceived during the training portion of the test. More specifically, this listener

- kept a written record of performance during training;
- was more succinct when describing in questionnaire answers ("a target echo usually does not contain a double echo"; "the clutter usually is either less crisp that target echoes or has a double sound");
- and, in one answer, the listener gives an example where it was better to rely on immediate experience gained during a current operation rather than past training experience gained ashore.

None of the results for the reduced-band test, including those associated with subject *s09,* are statistically significant.

3.3.1.4 Revised results of mean performance

A binormal ROC curve was deemed inappropriate for modelling six cases of individual subject performance. In four cases, no realistic binormal ROC curve could represent the measured data. In the two other cases, it was statistically proven, using a randomization test of the chi-square goodness-of-fit, that a binormal ROC curve could not represent the measured data.

Given that the model is inappropriate for these cases, it follows that any derived quantity, the numeric index A_z included, should not be used to represent performance. Revised human-performance results, with these six cases excluded, are presented in Table 6 and Table 7. The measures of average human performance that are included in the results are based only on the valid models. A comparison of results will be carried out in Section 3.5.

[5] It was also noted that subject *s03*'s approach to training for the Rating Exercises was different from other participants. In Training Exercise #2, where the listener was asked to practise making a classification decision, most participants would make use of the replay option both before and after the actual classification of the echo was presented to them. According to the session log, subject *s03* did not replay any of the sounds after the classifications had been revealed.

Table 6: Revised human-performance results for the full-band test. Values of A_z, standard deviations of A_z (σ_{Az}), and measures of goodness-of-fit (Q) are reported when the binormal ROC curve is deemed appropriate for modelling subject performance. These values are used in the calculation of average human performance (also shown). It is noted that the results for Rating Exercise #3 for subjects s03 and s09 are taken into account when calculating the mean value of A_z in the measure of average performance.

Listener code	Rating Exercise #1			Rating Exercise #3		
	A_z	σ_{Az}	Q	A_z	σ_{Az}	Q
s01	0.99	0.01	1.00	model not appropriate		
s03	model not appropriate			0.97	0.01	0.97
s04	0.99	0.01	0.89	0.99	0.01	0.99
s05	0.97	0.01	0.99	0.96	0.01	0.78
s06	model not appropriate			model not appropriate		
s07	0.77	0.04	1.00	0.82	0.04	0.43
s08	1.00	0.00	0.99	0.99	0.01	0.95
s09	model not appropriate			0.96	0.01	0.41
s10	0.98	0.01	0.94	0.98	0.01	1.00
s11	0.83	0.04	0.99	0.88	0.04	0.96
s12	0.98	0.01	0.94	0.97	0.01	0.62
s13	0.96	0.02	0.65	0.99	0.01	0.99
s14	0.97	0.01	0.59	0.94	0.02	0.79
s15	0.95	0.02	1.00	0.98	0.01	1.00
average performance	$A_{z\,mean}$ = **0.95**			$\sigma_{Az\,mean}$ = **0.02**		

Table 7: Revised human-performance results for the reduced-band test. Values of A_z, standard deviations of A_z (σ_{Az}), and measures of goodness-of-fit (Q) are reported when the binormal ROC curve is deemed appropriate for modelling subject performance. These values are used in the calculation of average human performance (also shown).

Listener code	Rating Exercise #1			Rating Exercise #3		
	A_z	σ_{Az}	Q	A_z	σ_{Az}	Q
s01	0.65	0.05	0.96	0.67	0.05	0.98
s04	0.70	0.04	0.99	0.71	0.04	0.88
s06	0.79	0.05	1.00	model not appropriate		
s08	0.91	0.03	1.00	0.90	0.03	1.00
s09	0.76	0.04	0.97	0.68	0.05	0.87
s10	0.69	0.04	0.73	0.75	0.04	0.95
s12	0.82	0.04	0.99	0.76	0.04	1.00
s15	0.75	0.04	0.54	0.73	0.04	0.96
s16	0.87	0.03	1.00	0.85	0.04	0.49
average performance	$A_{z\,mean}$ = **0.77**			$\sigma_{Az\,mean}$ = **0.04**		

3.3.2 Combining category frequencies to apply the chi-square test

In the second approach to assessing the gooness-of-fit, the measured data were tailored to suit the functionality of the analysis tool the reverse was true in the first approach. The idea was to combine category frequencies such that it might be reasonable to apply the chi-square test. For each subject and each Rating Exercise, adjacent rating categories were combined such that the smallest resulting category frequency was at least five. A record of the re-assembled input data can be found in Annex B, along with detailed results. The main trends and limitations of the results are discussed here.

The chief obstacle encountered when re-assembling the full-band data was the inability to combine the data yet still maintain at least three rating categories to produce a realistic binormal ROC curve. It was possible only for two subjects, and these were not among the better performers. Therefore a measure of average performance based only on these results considerably underestimates how well the group of subjects performed as a whole. This was less of a problem for the reduced-band data, where only one subject was not represented in the results.

Another difficulty was the potential unreliability of the chi-square test despite satisfying the criterion used in the web-based program to apply the test. The unreliability referred to here is not

the inaccuracy of the program's implementation of the chi-square test that was noted in Section 2.3.5; a separate chi-square test had been implemented in MATLAB® to address this shortcoming. The unreliability is due to the fact that the criterion is only a rule-of-thumb, and is not foolproof. The results, however, could be checked using the randomization test developed as the alternative to the chi-square test.

Eight of the 19 runs involved data lumped into four rating categories. (Many of the other runs were limited to three rating categories, for which a goodness-of-fit test was not necessary because the fit was perfect, and a small number of runs involved five rating categories, which the chi-square test could handle adequately.) Of the eight, one produced a measure of goodness-of-fit that was statistically significant. But the corresponding measure from the randomization test proved that it was indeed unreliable.

Another drawback of combining category frequencies is that information about the resolution with which the human listeners perceive differences is lost. Reducing the resolution, however, is exactly what is needed to treat the automatic classifier like a participant in the listening tests, and thus makes it possible to compare human and machine performances. The performance of the automatic classifier as a test participant is discussed in Section 3.4. Performance comparisons are left for Section 3.5.

3.4 The automatic classifier as listening-test participant

The automatic classifier that was developed during the course of the TIF project does not know how to interpret words such as "*is possibly CLUTTER*" or "*is definitely, or almost definitely, a TARGET echo*". But it does produce an index (a ratio of a posteriori probabilities), which is used to rank a series of echoes from most clutter-like to most target-like. If these ranks are binned, and if the number of target-echo cases and clutter cases in each bin are counted, then one has generated category frequencies, i.e. input data for this ROC analysis.

The details of how the automatic classifier works are outside the scope of this report; if interested, the reader can consult [2][3]. Given here is a summary of the steps that were taken to transform automatic classifier results into rating-scale data (i.e., treating the automatic classifier as if it were a participant in the full-band and reduced-band listening tests):

1. Use the training set of echoes in the listening test to train the automatic classifier and specify that it is to tailor its decision space so as to minimize the number of incorrect classification decisions.

2. Provide it with the testing echoes from the Rating Exercises and record the indices that it generates for these echoes.

3. Rank the echoes according to index.

4. Go through the list, forming bins that contain at least n target-echo cases and n clutter cases.

5. Report the number of target-echo cases and clutter cases in each bin as the category frequencies. A record of category frequencies representing the automatic classifier's performance in the listening tests is found in Annex C.

The bin width for creating equivalent rating categories was made somewhat flexible through the use of the parameter *n*. At first it was set to five, in order to observe results from the chi-square test (knowing that the goodness-of-fit could be validated by the randomization test). This setting led to only two bins (rating categories) being generated to represent classification performance of the full-band echoes, to which it was impossible to fit a realistic binormal ROC curve to model the performance. It is recalled that when trying to combine category frequencies of the human-performance data while specifying a minimum category frequency of five, a similar result was observed for the majority of test subjects. Realistic binormal ROC curves were obtained, however, when the parameter was set to a value less than five; all of these curves produced nearly identical values of A_z and σ_{Az}. When assembling the reduced-band data, the parameter n was kept at five because any smaller value produced more than seven bins (rating categories). Models of the automatic classifier's performance as a listening-test participant are summarized in Table 8.

Table 8: Performance results for the automatic classifier.

Automatic classifier	Rating Exercise		
	A_z	σ_{Az}	Q
full-band test	0.98	0.01	1.00
reduced-band test	0.86	0.03	1.00

3.5 Comparison of human and machine performance

Measures of average performance for each listening-test participant, including that of all the human listeners, are depicted in Figure 10 (in increasing order of A_{z_mean}). The measures of performance were obtained using the relationships described in Sections 2.4.2 and 2.4.3. Because the automatic classifier has no variability in its results given fixed input parameters, its measures of average performance simply correspond to the results reported in Table 8. It is also noted that for those listeners for whom a reliable model could be produced for only one Rating Exercise, their measures of average performance are based only on the one result. The error bounds depicted in Figure 10 are 95% confidence intervals (i.e. calculated as $\pm 1.96\sigma_{Az\_mean}$).

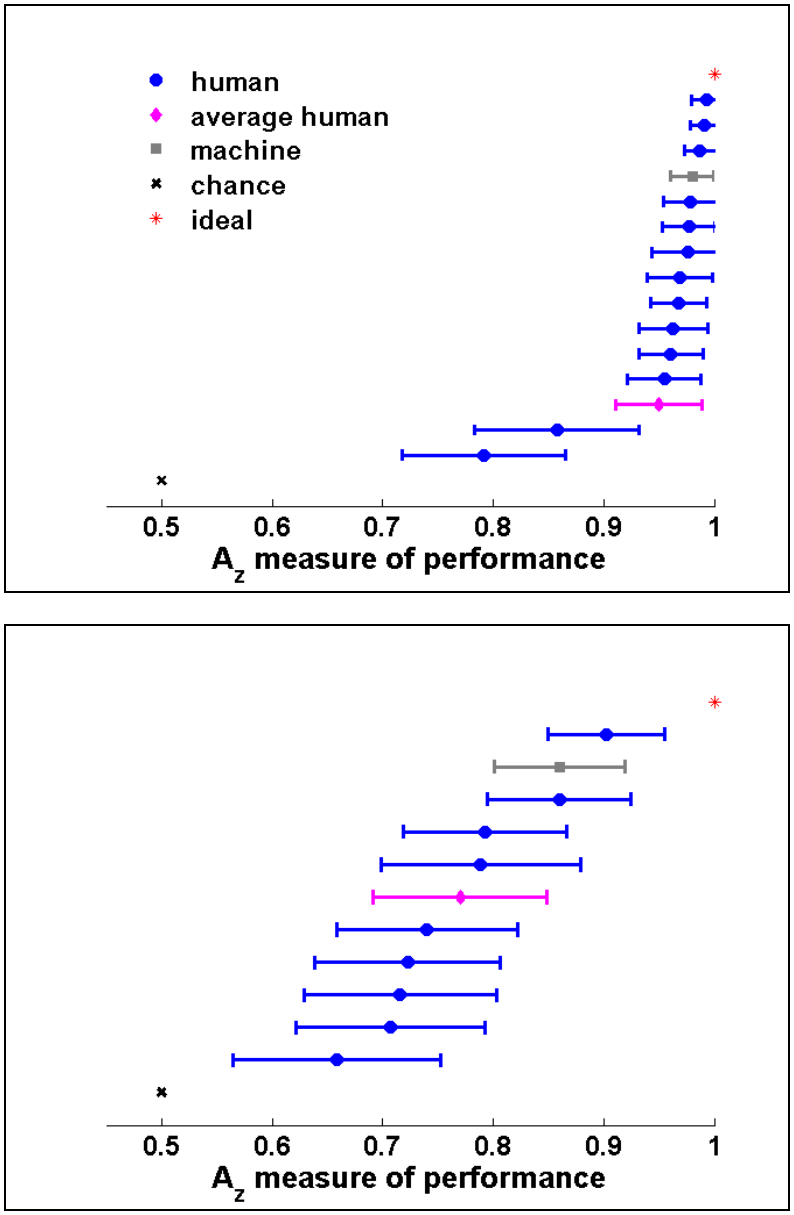

Figure 10: Measures of performance for the full-band test (top plot) and reduced-band test (bottom plot). Shown are mean values and 95% confidence intervals of A_z representing the performance of individual human listeners, average performance of all human listeners, and performance of the automatic classifier. Shown for reference are the theoretical values of A_z corresponding to chance performance and ideal performance.

A number of performance comparisons can be made by a simple visual inspection of Figure 10. A first obvious result is that all the full-band and reduced-band measures of performance are

significantly greater than 0.5, which corresponds to chance performance. The participants were therefore making use of aural differences in the target echoes and clutter when rating the echoes. Another is that, on the whole, performances in the full-band test are significantly better than those for the reduced-band test. One would conclude from this observation that aural cues are either less salient or absent when the information in the 0 □500 Hz band is removed. A number of the full-band results approach ideal performance ($A_z = 1$).[6]

Also noticeable is that the machine listener is significantly better than at least two of the human listeners in the full-band test and two in the reduced-band test. The statistical test described in Section 2.4.4 can be used to further qualify this observation. Results of this z-test are reported in Table 9 and Table 10. Significant results are obtained for two of the human listeners in the full-band test and five in the reduced-band test, indicating that the automatic classifier significantly outperformed these listeners. Other results are not significant: the widths of the confidence intervals are too large for the test to further discriminate between the performances. If these poorer performers are excluded from the calculation of average human performance (this re-calculation produces A_z = 0.97±0.02 for the full-band test and A_z = 0.85±0.08 for the reduced-band test), human performance closely resembles that of the automatic classifier (see Figure 11).

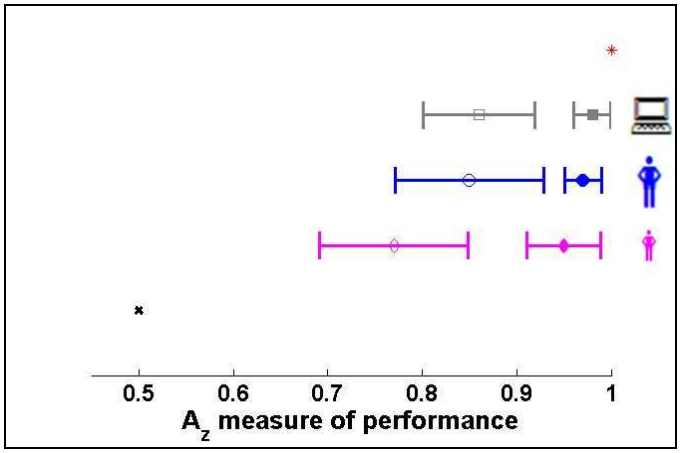

Figure 11: Plot summarizing the A_z average measures of performance achieved by the automatic classifier (shown in gray with the 💻 tag), a subset of the better human listeners (shown in blue with the 🧍 tag), and all the human listeners (shown in magenta with the 🧍 tag). Chance and ideal performances are also depicted. The solid symbols (■,●,♦) and associated error bounds are the full-band results; the hollow symbols and associated error bounds are the reduced-band results. It was statistically proven that the performance of the automatic classifier was better than 2 of the 13 listeners in the full-band test and 5 of the 9 listeners in the reduced-band test (the subset of better human listeners excludes these poorer performers).

[6] The confidence interval in some cases slightly extends past 1: this error arises because the distribution of A_z becomes skewed at very high values. The exact numbers of the high-performance results should therefore be used with caution. But the trend of very good performance in the full-band tests remains valid.

Table 9: Using a z-test to compare the performances of the automatic classifier and the human listeners in the full-band test. Results show that the automatic classifier was significantly better than two of the 13 human listeners (using a 5% significance level).

Listener code	z-statistic	same (□) / better (o) / worse (x)
s01	0.87	□
s03	+0.71	□
s04	0.49	□
s05	-0.85	□
s07	-4.91	x
s08	1.05	□
s09	-1.16	□
s10	-0.27	□
s11	-3.16	x
s12	-0.22	□
s13	-0.25	□
s14	-1.37	□
s15	-1.01	□
average	-1.38	□

Table 10: Using a z-test to compare the performances of the automatic classifier and the human listeners in the reduced-band test. Results show that the automatic classifier was significantly better than five of the nine human listeners (using a 5% significance level).

Listener code	z-statistic	same (□) / better (o) / worse (x)
s01	-3.40	x
s04	-2.73	x
s06	-1.20	□
s08	1.05	□
s09	-2.53	x
s10	-2.48	x
s12	-1.29	□
s15	-2.19	x
s16	0.05	□
average	-1.59	□

3.6 Usefulness of qualitative data

3.6.1 Sample size

The standard error in the measures of performance was determined largely by the sample variance. If the size of the sample of echoes had been larger, the sample variance would have decreased and, consequently, so would the standard error. A decrease in the standard error means a more accurate estimate of the measure of performance. It also means an improved "resolution" of the statistical test: smaller differences between two measures of performance can be proven to be significantly different.

But the larger the sample, the more likely factors such as fatigue and stress would influence human-performance results. One of the participants in fact reported feeling quite tired after the listening-test session. Another reported feeling tired going into the test and having difficulty concentrating on the sounds. A small number of subjects reported that, the more they listened to echoes, the more they started doubting and questioning their ability to classify them. It would not be unreasonable to think that these test subjects, consciously or unconsciously, modified their rating criteria during the course of the test session. Others seemed to consider the listening test a race and were inclined to rush decisions in order to beat some imaginary stopwatch. These comments from the test participants were a useful reminder that the simple "remedy" of increasing sample size to increase the accuracy of the measure of performance would also likely introduce undesirable side-effects that would adversely affect human performance.

3.6.2 Understanding odd results

In a small number of cases, a value of A_z could not be produced because no ROC curve could be fitted to the data. This occurred because the listener used only two rating categories. It could be inferred from qualitative data collected during the tests that the reason for this pattern was not because of a clear-cut distinction in aural cues but because of the listener's distinct preference for a black-and-white decision-making strategy: either it is a target echo or it is not. One listener reported using the same strategy in military sonar operations. Although no claim is made here that it is incorrect, the strategy could not, given the design of the listening test, produce valid data for analysis. These data were therefore disregarded.

Another small set of data was disregarded because statistical testing proved that the binormal ROC curve was inappropriate for modelling the data. Consulting once again the qualitative data to explain the odd result, the likely reason seemed to be that, because of their training and extensive use of Doppler in passive sonar operations, the listeners were trying to use it as a cue for making their decisions, which would be very misleading for classifying the active sonar echoes in the listening tests. Therefore these data are also not represented in the performance results. The finding also suggests that the information provided to prospective participants could be improved – something to bear in mind when designing possible future tests.

4 Conclusion

One of the aims of the listening tests was to corroborate anecdotal evidence that sonar operators could hear differences between target echoes and clutter. The performance results from the tests, graphically summarized in Figure 10 and Figure 11, strongly support the anecdotal evidence: the measures of performance A_z of all the human subjects for both the full-band and reduced-band tests are significantly higher than the value of 0.5 that corresponds to chance performance (which would reflect an inability to hear differences in the echoes).

Another aim of the listening tests was to provide a baseline measure of performance against which the performance of the automatic classifier could be compared. Although the automatic classifier did not actually "participate" in the listening tests, it was possible to treat the results it generates in a way that was similar to listening-test data (see Figure 10 and Figure 11). A one-on-one performance comparison was carried out between the "machine" participant and each of the human participants of the listening tests (it is recalled that the human listeners were all DND personnel with significant experience in sonar). Results of a statistical test proved that the automatic classifier outperformed at least two of the 13 listeners in the full-band test and five of nine listeners in the reduced-band test and was on par with the other listeners – further testing and analysis would be required to distinguish between the performances.

Finally, a comparison was made between the full-band and reduced-band performances. Statistical testing showed that removing echo content in the band 0 – 500 Hz led to a significant performance degradation for both the human and machine listeners.

References

[1] Hines, P.C. (2002). Aural Discrimination of True Targets from Geological Clutter, *DRDC* Technology *Investment Fund Project Proposal,* DRDC Atlantic.

[2] Young, W.V. (2005). Application of Musical Timbre Discrimination Features to Active Sonar Classification. MSc thesis, Dalhousie University, Halifax, Nova Scotia, Canada.

[3] Young, V.W. and Hines, P.C. (2007). Perception-based automatic classification of impulsive-source active sonar echoes. *J. Acoust. Soc. Am,* Vol.122, No.3, pp.1502 1517.

[4] Allen, N. and Hines, P.C. (2007). Protocols for two human performance tests on aural discrimination of sonar echoes. (DRDC Atlantic TM 2006-302). Defence R&D Canada Atlantic.

[5] Green, D.M. and Swets, J.A. (1988). Signal Detection Theory and Psychophysics. Los Altos, California: Peninsula Publishing.

[6] Abramowitz, M. and Stegun, I.A. (ed.) (1965). Handbook of Mathematical Functions. NewYork: Dover Publications.

[7] Swets, J.A. and Pickett, R.M. (1982). Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. New York: Academic Press.

[8] Swets, J.A. (ed.) (1964). Signal Detection and Recognition by Human Observers. NewYork: John Wiley & Sons.

[9] Eng, J. ROC analysis: web-based calculator for ROC curves (online). Johns Hopkins University, http://www.jrocfit.org (12 September 2007).

[10] Dorfman, D.D. and Alf, Jr. E. (1969). Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals Rating-Method Data. *Journal of Mathematical Psychology,* Vol.6, pp.487 496.

[11] Grey, D.R. and Morgan, B.J.T. (1972). Some Aspects of ROC Curve-Fitting: Normal and Logistic Models. *Journal of Mathematical Psychology,* Vol.9, pp.128 139.

[12] Metz, C.E., Shen, J.-H., Wang, P.-L., and Kronman, H.B. (1994). ROCFIT program in Fortran (modified version of the program RSCORE II by Donald D. Dorfman). University of Chicago, http://www.bio.ri.ccf.org/doc/rocfit.f (12 September 2007).

[13] Au, W.W.L. and Martin, D.W. (1989). Insights into dolphin sonar discrimination capabilities from human listening experiments. *Journal of the Acoustical Society of America,* Vol.86, No.5, pp.1662 1670.

[14] McFadden, S.M. (1984). Auditory Display of Passive Sonar Information. (DCIEM No. 84-R-23). Defence and Civil Institute of Environmental Medicine.

[15] Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992). Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd edition. Cambridge University Press.

[16] Lapin, L.L (1980). Statistics: Meaning and Method, 2nd edition. New York: Harcourt Brace Jovanovich, Inc.

[17] Walpole, R.E. and Myers, R.H. (1978). Probability and Statistics for Engineers and Scientists, 2nd edition. New York: MacMillan Publishing Co., Inc.

[18] Davies, O.L. (ed.) (1949). Statistical Methods in Research and Production, second edition. Edinburgh: Oliver and Boyd.

[19] Mead, R. (1988). The design of experiments: Statistical principles for practical application. Cambridge: Cambridge University Press.

This page intentionally left blank.

Annex A ROC input data from the listening tests

Table 12 – Table 17 contain the category frequencies (term first explained in Section 2.2.6) that were obtained from the human subjects' responses to Rating Exercise #1 and Rating Exercise #3 in the full-band and reduced-band listening tests. Please note that data from participant *s02* do not appear in the tables: because a different set of stimuli was inadvertently used during the full-band and reduced-band tests for participant *s02* (all other participants were presented with the same stimuli) and given the good turnout of volunteers for the tests, it seemed reasonable to discard responses to the stray data set of stimuli.

Table 11: Category frequencies □Clutter cases □Full-band test □Rating Exercise #1

Rating scale / Listener code	definitely clutter	probably clutter	possibly clutter	don't know	possibly a target echo	probably a target echo	definitely a target echo
s01	37	16	2	3	6	8	1
s03	17	32	1	8	5	9	1
s04	50	17	0	2	2	2	0
s05	56	8	1	4	1	1	2
s06	52	4	6	1	0	3	7
s07	21	17	12	10	2	7	4
s08	58	13	1	1	0	0	0
s09	13	24	15	18	3	0	0
s10	28	29	0	8	8	0	0
s11	37	18	0	1	0	9	8
s12	46	16	2	6	3	0	0
s13	48	21	0	4	0	0	0
s14	4	57	0	7	5	0	0
s15	45	21	2	3	1	1	0

Table 12: Category frequencies □Target-echo cases □Full-band test □Rating Exercise #1

Rating scale / Listener code	definitely clutter	probably clutter	possibly clutter	don't know	possibly a target echo	probably a target echo	definitely a target echo
s01	0	0	0	0	1	9	63
s03	1	0	0	0	21	40	11
s04	0	2	2	0	6	13	50
s05	2	3	0	1	1	9	57
s06	2	0	0	0	0	0	71
s07	6	8	6	10	6	10	27
s08	0	1	3	1	8	19	41
s09	0	1	9	1	21	23	18
s10	0	1	0	5	12	28	27
s11	2	10	0	1	3	30	27
s12	1	2	1	0	3	27	39
s13	3	1	1	1	9	33	25
s14	0	1	3	12	20	25	12
s15	3	8	1	4	4	13	40

Table 13: Category frequencies □Clutter cases □Full-band test □Rating Exercise #3

Rating scale / Listener code	definitely clutter	probably clutter	possibly clutter	don't know	possibly a target echo	probably a target echo	definitely a target echo
s01	35	13	6	0	5	11	3
s03	11	27	2	21	10	2	0
s04	42	15	1	3	8	4	0
s05	54	6	2	3	1	3	4
s06	51	0	6	0	0	0	16
s07	15	20	3	18	8	6	3
s08	53	12	1	6	1	0	0
s09	22	20	13	6	9	2	1
s10	23	33	0	13	4	0	0
s11	39	17	0	0	0	6	11
s12	46	18	0	6	2	1	0
s13	42	26	2	3	0	0	0
s14	1	53	1	15	3	0	0
s15	61	9	1	1	1	0	0

Table 14: Category frequencies □Target-echo cases □Full-band test □Rating Exercise #3

Rating scale / Listener code	definitely clutter	probably clutter	possibly clutter	don't know	possibly a target echo	probably a target echo	definitely a target echo
s01	0	0	0	0	0	2	71
s03	1	1	0	0	8	58	5
s04	0	1	1	0	3	18	50
s05	3	0	0	0	6	8	56
s06	0	0	0	0	0	0	73
s07	0	10	6	7	7	10	33
s08	1	0	1	2	2	17	50
s09	0	2	1	1	6	43	20
s10	1	1	0	3	18	28	22
s11	0	3	0	0	2	13	55
s12	1	1	4	5	6	15	41
s13	0	3	0	3	11	34	22
s14	0	4	3	10	31	22	3
s15	1	12	6	6	7	24	17

Table 15: Category frequencies □Clutter cases □Reduced-band test □Rating Exercise #1

Rating scale / Listener code	definitely clutter	probably clutter	possibly clutter	don't know	possibly a target echo	probably a target echo	definitely a target echo
s01	0	2	8	41	19	3	0
s04	7	21	10	15	17	3	0
s06	50	0	3	0	0	0	20
s08	37	14	3	5	4	4	6
s09	36	14	5	1	0	12	5
s10	12	13	5	20	21	2	0
s12	34	19	3	4	5	8	0
s15	23	23	4	6	15	2	0
s16	48	8	3	0	8	4	2

Table 16: Category frequencies □Target-echo cases □Reduced-band test □Rating Exercise #1

Rating scale / Listener code	definitely clutter	probably clutter	possibly clutter	don't know	possibly a target echo	probably a target echo	definitely a target echo
s01	0	1	4	26	37	5	0
s04	5	7	9	7	28	12	5
s06	19	0	2	0	0	0	52
s08	2	2	1	3	6	16	43
s09	12	14	1	1	0	19	26
s10	1	7	10	14	31	10	0
s12	7	10	4	9	3	27	13
s15	1	21	10	12	15	13	1
s16	2	8	7	4	20	24	8

Table 17: Category frequencies □Clutter cases □Reduced-band test □Rating Exercise #3

Rating scale / Listener code	definitely clutter	probably clutter	Possibly clutter	don't know	possibly a target echo	probably a target echo	definitely a target echo
s01	0	0	11	41	20	1	0
s04	2	26	10	11	21	3	0
s06	55	0	2	0	0	0	16
s08	38	12	2	6	5	3	7
s09	36	14	0	0	0	14	9
s10	9	16	12	16	20	0	0
s12	16	24	2	14	10	4	3
s15	24	23	5	6	9	6	0
s16	41	7	5	1	2	14	3

Table 18: Category frequencies □Target-echo cases □Reduced-band test □Rating Exercise #3

Rating scale / Listener code	definitely clutter	probably clutter	possibly clutter	don't know	possibly a target echo	probably a target echo	definitely a target echo
s01	0	1	4	28	34	6	0
s04	1	6	6	19	27	11	3
s06	20	0	0	0	0	0	53
s08	1	3	2	4	6	13	44
s09	16	19	3	0	0	15	20
s10	0	6	9	15	33	10	0
s12	2	10	3	10	22	17	9
s15	6	20	1	5	20	17	4
s16	0	3	6	5	16	29	14

Annex B Re-assembled ROC input data and results

Table 19 to Table 26 contain the re-assembled ROC input data for the analysis described in Section 3.3.2. The category frequencies in Annex A were combined such that the minimum observed category frequency is five. This combination of numbers from adjacent rating categories was done individually for each set of input data (the category frequencies for the target-echo cases and clutter cases for one Rating Exercise, for one human listener, and for one test). The five general labels 'A', 'B', 'C', 'D', 'E' are used in the tables to designate the newly formed rating categories, since five was the maximum number of categories encountered during the re-assembly. For cases that led to three or four categories, the symbol '□' is used in the tables to designate the unused categories. Those cases that produced less than three categories are not displayed because the data does not produce a realistic binormal ROC curve (see Section 3.2).

Also included here, in Table 27 to Table 30, are the modelling results, including a comparison of goodness-of-fit results from the chi-square test and the randomization test (described in Section 3.3.1).

Table 19: Combined category frequencies □Clutter cases □Full-band test □Rating Exercise #1

Rating scale / Listener code	A	B	C	D	E
s07	21	17	12	10	13
s11	55	10	8	□	□

Table 20: Combined category frequencies □Target-echo cases □Full-band test □ Rating Exercise #1

Rating scale / Listener code	A	B	C	D	E
s07	6	8	6	10	43
s11	12	34	27	□	□

Table 21: Combined category frequencies □Clutter cases □Full-band test □Rating Exercise #3

Rating scale / Listener code	A	B	C	D	E
s07	25	21	8	9	□
s11	56	6	11	□	□

Table 22: Combined category frequencies □Target-echo cases □Full-band test □ Rating Exercise #3

Rating scale / Listener code	A	B	C	D	E
s07	10	13	7	43	□
s11	5	13	55	□	□

Table 23: Combined category frequencies □Clutter cases □Reduced-band test □ Rating Exercise #1

Rating scale / Listener code	A	B	C	D	E
s01	10	41	22	□	□
s04	7	21	10	15	20
s08	54	13	6	□	□
s09	36	14	18	5	□
s10	25	5	20	23	□
s12	34	19	12	8	□
s15	46	10	17	□	□
s16	56	11	6	□	□

Table 24: Combined category frequencies □Target-echo cases □Reduced-band test □ Rating Exercise #1

Rating scale / Listener code	A	B	C	D	E
s01	5	26	42	□	□
s04	5	7	9	7	45
s08	5	25	43	□	□
s09	12	14	21	26	□
s10	8	10	14	41	□
s12	7	10	16	40	□
s15	22	22	29	□	□
s16	10	31	32	□	□

Table 25: Combined category frequencies □Clutter cases □Reduced-band test □ Rating Exercise #3

Rating scale / Listener code	A	B	C	D	E
s01	11	41	21	□	□
s04	28	10	11	24	□
s08	52	11	10	□	□
s09	36	14	14	9	□
s10	25	12	16	20	□
s12	40	16	10	7	□
s15	24	23	11	9	6

Table 26: Combined category frequencies □Target-echo cases □Reduced-band test □ Rating Exercise #3

Rating scale / Listener code	A	B	C	D	E
s01	5	28	40	□	□
s04	7	6	19	41	□
s08	6	10	57	□	□
s09	16	19	18	20	□
s10	6	9	15	43	□
s12	12	13	22	26	□
s15	6	20	6	20	21

Table 27: Results for the full-band test following the approach of combining category frequencies to apply the chi-square goodness-of-fit test. Values of A_z, standard deviations of A_z (σ_{Az}), and probabilities from the chi-square test (Q-X^2) are shown. Average performance is also shown.

Listener code	Rating Exercise #1			Rating Exercise #3		
	A_z	σ_{Az}	Q-X^2	A_z	σ_{Az}	Q-X^2
s07	0.77	0.04	0.88	0.78	0.04	0.72
s11	0.84	0.04	□	0.89	0.04	
average performance	**0.81**	**0.04**				

□Perfect fit. The input data provided only two operating points, to which a binormal ROC curve could be fit exactly. A goodness-of-fit test was therefore not needed.

Table 28: Comparison of results from the chi-square test and the randomization test for the full-band models: X^2 is the observed test statistic, dof is the number of degrees of freedom, Q-X^2 is the measure of goodness-of-fit from the chi-square test, and Q is the measure of goodness-of-fit from the randomization test.

Listener code	Rating Exercise #1				Rating Exercise #3			
	X^2	*dof*	Q-X^2	*Q*	X^2	*dof*	Q-X^2	*Q*
s07	0.27	2	0.88	1.00	0.13	1	0.72	1.00
s11			□	□				

□Perfect fit. The input data provided only two operating points, to which a binormal ROC curve could be fit exactly. A goodness-of-fit test was therefore not needed.

Table 29: Results for the reduced-band test following the approach of combining category frequencies to apply the chi-square goodness-of-fit test. Values of A_z, standard deviations of A_z (σ_{Az}), and probabilities from the chi-square test ($Q\text{-}X^2$) are shown. Average performance is also shown. One result is found to be statistically significant, but unreliable (see Table 30 and Section 3.3.2).

Listener code	Rating Exercise #1			Rating Exercise #3		
	A_z	σ_{Az}	$Q\text{-}X^2$	A_z	σ_{Az}	$Q\text{-}X^2$
s01	0.68	0.05	[1]	0.68	0.05	
s04	0.72	0.05	0.27	0.68	0.05	0.36
s08	0.91	0.03		0.91	0.03	
s09	0.75	0.04	0.29	0.68	0.05	0.40
s10	0.67	0.05	**0.02**	0.72	0.05	0.62
s12	0.82	0.04	0.78	0.78	0.04	0.49
s15	0.70	0.05		0.73	0.04	0.20
s16	0.88	0.03		--[2]	--	**--**
average performance	**0.77**	**0.05**				

[1] Perfect fit. The input data provided only two operating points, to which a binormal ROC curve could be fit exactly. A goodness-of-fit test was therefore not needed.

[2] No model fitted to data – see footnote in Table 2.

Table 30: Comparison of results from the chi-square test and the randomization test for the reduced-band models: X^2 is the observed test statistic, dof is the number of degrees of freedom, Q- X^2 is the measure of goodness-of-fit from the chi-square test, and Q is the measure of goodness-of-fit from the randomization test. The randomization test seems to confirm that the significant result of the chi-square test was in fact due to the unreliability of the test when there is only one degree of freedom and the minimum category frequency is less than 10.

Listener code	Rating Exercise #1				Rating Exercise #3			
	X^2	*dof*	Q- X^2	*Q*	X^2	*dof*	Q- X^2	*Q*
s01			□	□				
s04	2.64	2	0.27	0.96	0.85	1	0.36	0.99
s08	0				0			
s09	1.13	1	0.29	0.98	0.72	1	0.40	0.99
s10	5.07	1	**0.02**	0.54	0.24	1	0.62	1.00
s12	0.08	1	0.78	1.00	0.49	1	0.49	1.00
s15	0				3.27	2	0.20	0.91
s16	0				--□	--	**--**	--

□Perfect fit. The input data provided only two operating points, to which a binormal ROC curve could be fit exactly. A goodness-of-fit test was therefore not needed.

□No model fitted to data□ see footnote in Table 2.

Annex C ROC input data from the automatic classifier

Table 31 to Table 33 show category frequencies representing the automatic classifier's performance as listening-test participant. A description of how results from the automatic classifier were handled so as to represent rating-scale data is given in Section 3.4.

Table 31: Category frequencies representing the automatic classifier's performance in the full-band test. When specifying a minimum category frequency of five, only two rating categories (bins) are produced.

Rating scale / Class of echo	A	B
Target echoes	5	68
Clutter	67	6

Table 32: Category frequencies representing the automatic classifier's performance in the full-band test. When specifying a minimum category frequency of three, four rating categories (bins) are produced, with the fourth (the "A" rating category) containing the remaining echoes that could not form a complete bin.

Rating scale / Class of echo	A	B	C	D
Target echoes	2	3	3	65
Clutter	62	5	3	3

Table 33: Category frequencies representing the automatic classifier's performance in the reduced-band test. When specifying a minimum category frequency of six, five rating categories (bins) are produced, with the fifth (the "A" rating category) containing the remaining echoes that could not form a complete bin.

Rating scale / Class of echo	A	B	C	D	E
Target echoes	49	6	6	6	6
Clutter	5	6	14	22	26

This page intentionally left blank.

Distribution list

Document No.: DRDC Atlantic TM 2007-353

LIST PART 1: Internal Distribution by Centre:

1 Senior Military Officer
1 H/US
2 P.C. Hines (1 hardcopy + 1 softcopy)
2 N. Allen (1 hardcopy + 1 softcopy)
5 Library

11 TOTAL LIST PART 1

LIST PART 2: External Distribution by DRDKIM
Department of National Defence

1 DRDKIM

1 ADAC(A) Commanding Officer
1 CFMWC Commanding Officer
1 CFNOS Commanding Officer
1 MOG5 Commanding Officer
1 PSU HALIFAX Commanding Officer
1 TRINITY Commanding Officer
address for these 6 DND units: PO Box 99000 Stn Forces, Halifax, NS B3K 5X5

International

1 Dr. James Pitton
ONR Global
Edison House
223 Old Marylebone Rd.
London NW1 5TH
UK

1 Dr. James A. Ballas
Naval Research Laboratory
Code 5513
Washington, DC 20375-5337
USA

1 Duncan Williams
Defence Science & Technology Laboratory, Physical Sciences
Winfrith Technology Centre
Dorchester, Dorset DT2 8XJ
UK

continued on following page

LIST PART 2: External Distribution by DRDKIM (continued)
International (continued)

1 Dr. Tor Knudsen
Director of Research, FFI
PO Box 115
N3191, Horten
Norway

1 Dr. David Liebing
National Leader, TTCP MAR TP-9
Defence Science & Technology Organisation, Maritime Operations Division
PO Box 1500
Edinburgh SA 5111
AUSTRALIA

1 Mr. Matthew Hopkins
National Leader, TTCP MAR TP-9
Defence Technology Agency
Naval Base, Private Bag 32901
Devonport, Auckland
NEW ZEALAND

1 Dr. Geoffrey A. Williams
National Leader, TTCP MAR TP-9
Defence Science & Technology Laboratory, Winfrith Technology Centre
Winfrith Newburgh
Dorchester, Dorset DT2 8XJ
UK

1 Dr. Edward R. Franchi
National Leader, TTCP MAR TP-9
Naval Research Laboratory
4555 Overlook Avenue SW
Washington DC 20375
USA

1 Dr. Keith L. Davidson
Undersea Signal Processing (321US)
Office of Naval Research
875 North Randolph Street, Suite 1425
Arlington, VA 22203-1995
USA

16 TOTAL LIST PART 2

27 TOTAL COPIES REQUIRED

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.)

Defence R&D Canada – Atlantic
9 Grove Street
P.O. Box 1012
Dartmouth, Nova Scotia B2Y 3Z7

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

unclassified

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C, R or U) in parentheses after the title.)

Receiver-Operating-Characteristic (ROC) analysis applied to listening-test data: Measures of performance in aural classification of sonar echoes

4. AUTHORS (last name, followed by initials □ranks, titles, etc. not to be used)

Allen, N.

5. DATE OF PUBLICATION (Month and year of publication of document.)	6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)	6b. NO. OF REFS (Total cited in document.)
August 2008	74	19

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Technical Memorandum

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development □include address.)

Defence R&D Canada – Atlantic
9 Grove Street
P.O. Box 1012
Dartmouth, Nova Scotia B2Y 3Z7

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)
11cq11	

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC Atlantic TM 2007-353	

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

(X) Unlimited distribution
( ) Defence departments and defence contractors; further distribution only as approved
( ) Defence departments and Canadian defence contractors; further distribution only as approved
( ) Government departments and agencies; further distribution only as approved
( ) Defence departments; further distribution only as approved
( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

A Receiver-Operating-Characteristic analysis was conducted on data from two listening tests that were carried out as part of a DRDC Technology Investment Fund project on aural classification of sonar echoes. All the human listeners were DND personnel with significant experience in sonar. An automatic classifier was also tested. The results of the analysis strongly support the idea of using aural cues to discriminate between target echoes and clutter. The automatic classifier outperformed some of the human listeners and was on par with the others. Performances of the human listeners and automatic classifier in the second listening test, where a high-pass filter was applied to the echoes to remove all frequency content below 500 Hz, were definitely poorer than for the first test, where the full available bandwidth of 0 □2 kHz was exploited, but still substantially better than chance performance.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

measures of performance, human-performance testing, ROC analysis, aural classification, sonar-echo classification

This page intentionally left blank.

Defence R&D Canada

Canada's leader in defence
and National Security
Science and Technology

R & D pour la défense Canada

Chef de file au Canada en matière
de science et de technologie pour
la défense et la sécurité nationale

DEFENCE DÉFENSE

www.drdc-rddc.gc.ca