



Queensland University of Technology
Brisbane Australia

This is the author's version of a work that was submitted/accepted for publication in the following source:

Truskinger, Anthony Masters, Yang, Haofan, Wimmer, Jason, Zhang, Jinglan, Williamson, Ian, & Roe, Paul (2011) Large scale participatory acoustic sensor data analysis : tools and reputation models to enhance effectiveness. In Werner, Bob (Ed.) *7th IEEE International Conference on eScience*, IEEE Computer Society, Stockholm City Conference Centre/Norra Latin, Stockholm. (In Press)

This file was downloaded from: <http://eprints.qut.edu.au/45996/>

© Copyright 2011 IEEE

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Notice: *Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source:*

Large Scale Participatory Acoustic Sensor Data Analysis: Tools and Reputation Models to Enhance Effectiveness

Anthony Truskinger, Haofan Yang, Jason Wimmer, Jinglan Zhang, Ian Williamson and Paul Roe

Microsoft QUT eResearch Centre
Queensland University of Technology
Brisbane, Australia

(anthony.truskinger; haofan.yang)@student.qut.edu.au
(j.wimmer; jinglan.zhang; i.williamson; p.roe)@qut.edu.au

Abstract— Acoustic sensors play an important role in augmenting the traditional biodiversity monitoring activities carried out by ecologists and conservation biologists. With this ability however comes the burden of analysing large volumes of complex acoustic data. Given the complexity of acoustic sensor data, fully automated analysis for a wide range of species is still a significant challenge. This research investigates the use of citizen scientists to analyse large volumes of environmental acoustic data in order to identify bird species. Specifically, it investigates ways in which the efficiency of a user can be improved through the use of species identification tools and the use of reputation models to predict the accuracy of users with unknown skill levels. Initial experimental results are reported.

Keywords- *global climate change; sensors; acoustic sensing; acoustic analysis; citizen science; reputation management; participatory sensing; participatory analysis.*

I. INTRODUCTION

The complex and interrelated environmental effects of climate change present scientists with multi-faceted problems requiring innovative solutions, if we are to preserve biodiversity. With estimates of extinction rates up to 1000 times the natural rate, monitoring the effects of climate change on the earth's biodiversity is becoming increasingly important [1]. Conducting biodiversity monitoring at large spatiotemporal scales using traditional manual methods is costly, time-consuming and ultimately fails to provide scientists with the large scale, timely observations they require [2]. Acoustic sensors have the potential to play an important role augmenting the traditional biodiversity monitoring activities carried out by ecologists and conservation biologists [3-5]. They can be deployed for extended periods of time, across large areas, continuously and objectively recording the sounds of the environment. These sounds can then be analysed to determine vocal species which are present in the recordings.

The use of acoustic sensors for large scale spatiotemporal ecological research is an attractive proposition for many scientists. Sensors can remain deployed in the field across large areas, in remote locations for extended periods of time, and at a fraction of the cost of deploying human observers [6, 7]. With this ability however comes the burden of analysing large volumes of complex acoustic data [8]. Acoustic data recorded in the field is subject to extraneous environmental noise such as wind and rain; additionally many species also

demonstrate a vast repertoire of vocalisations, regional variation and even mimicry [9]. The raw sensor data must ultimately be filtered, analysed and processed to provide scientists with the species, population, and distribution information they require. Much eScience research has and is currently being done into automated approaches to processing acoustic sensor data, however given the complexity of acoustic sensor data, fully automated analysis for a wide range of species is still a significant challenge [8, 10-20].

Participatory data analysis uses citizen scientists to analyse large volumes of data. It offers a potential solution for analysing large volumes of sensor data and is particularly good at solving eScience problems [21-23]. Participatory data analysis draws on the resources of volunteers and enthusiasts to manually analyse large volumes of complex data that may be difficult to analyse computationally [22, 23].

The inherent complexity of acoustic sensor data analysis lends itself to participatory analysis approaches which can take advantage of large numbers of participants who can collectively analyse large volumes of data. As in other citizen science projects however, the credibility of participant contributions must be established to ensure high levels of accuracy in analysis. This research investigates the use of citizen scientists to analyse large volumes of environmental acoustic data in order to identify bird species. Specifically it investigates the application of reputation models to predict the species identification accuracy of users analysing acoustic sensor data using our online acoustic workbench (<http://sensor.mquter.qut.edu.au>). This workbench has been developed to provide a comprehensive suite of acoustic sensor data analysis tools which users can access through a standard web browser. This also allows large numbers of citizen scientists to access and analyse large volumes of acoustic sensor data remotely. Common challenges often associated with participatory projects like this include: hard to use systems, inefficient analysis methods, and data reliability.

In this paper we suggest metrics for calculating reputation scores for acoustic sensor data analysis and techniques for using these metrics to rank potential participants based on past performance and initial trust. We demonstrate that determining the analysis reputation of participants improves data analysis quality and ensures that large-scale participatory acoustic data analysis can be reliable. This paper also demonstrates that by using simple methods of classification to help rank possible identification results, participant's efficiency and annotation

accuracy can be significantly improved. Together, these findings provide a promising basis for further participatory acoustic sensor data analysis work to build upon.

The remainder of this paper is organised as follows. Section II reviews related work. Section III presents the design of the software system and section IV reports the initial evaluation results. Section V discusses issues and section VI draws the conclusion and future work.

II. RELATED WORK

In many citizen science projects, participants contribute both by analyzing data (Galaxy Zoo: <http://www.galaxyzoo.org>) and collecting and contributing data (eBird: <http://www.ebird.org>). Given the varied background of citizen science participants (ranging from amateur enthusiasts to experienced scientists), there are a number of significant challenges to be overcome with citizen science projects [24]. One of the foremost challenges is establishing the skill level or reputation of the participant performing the collection or analysis task. To achieve this, many citizen science projects utilise reputation management to classify participants and to establish the credibility of their contributions.

Galaxy Zoo is a classic example of this approach, with over 250,000 active users helping to manually classify galaxy types according to their shapes [23]. Galaxy Zoo provides users with initial identification training and testing and then provides an interface for classifying galaxies, deferring the final complex analysis task to humans. Identification of the same galaxy by multiple users ensures consistency and accuracy. Since the data of citizen science projects is contributed by volunteers and most of them have little or even no scientific training the quality of contributed data is not guaranteed. To overcome this, some citizen science projects apply the concept of reputation management to classify contributors and use the results of subsequent human analysis tasks to assess the credibility of contributors [21, 25-27].

The Amazon Mechanical Turk (<http://www.mturk.com/>), a crowdsourcing project, coordinates the demand and supply of tasks that require human intelligence and creativity skills to complete. It provides a reputation mechanism to support the quality and credibility issues. The similarities between this mechanism and our research are that both use reputation-based approaches to deal with the trust and take the entity's past performance as the major element while modeling the reputation. However, the relationships between the participants and their taskers in Mechanical Turk is very different from our participatory sensing project. The Mechanical Turk crowdsources tasks from many taskers to many participants. Whereas our acoustic sensors project (and Galaxy Zoo to) crowdsources one large task to many participants. Thus even though both projects utilise crowdsourcing, trust reputation methodologies are not directly compatible.

Information technology has an important role to play in assisting and improving manual biodiversity surveys. Traditionally, identification of species in the field has been achieved with the assistance of field guides. Usually in book

form, these field guides contain the descriptions of many species, typically over large geographical areas. They often also have dedicated keys that help to improve the speed and accuracy of species identification in the field, although these keys usually require some level of existing knowledge to use effectively. With the advent of modern technology it has now become possible to store these field guides digitally.

Carrying a physical field guide has evolved into carrying a guide on a laptop and recently, into carrying a guide on a smart phone. Often these digital guides also include recorded examples of species vocalisations – a powerful innovation made possible through the widespread adoption of modern information communication technology. However, it is worth noting that because these guides are often produced in an ideal environment, they are generally not accurate representations of real world species vocalisations [28].

III. METHOD

A. Reputation Model

In many reputation systems, reputation models are populated using one of three methods: past performance of the targeted participant, the opinions of other users, or a combination of both. Depending on the information provided, the procedure for implementing the reputation model will differ. Wang and Zhang [29] state that “Trust is mainly a social phenomenon”, that is, any reputation model should be based on how trust works in society. An example of a real life, well-known, reputation system is eBay; eBay supports transaction records directly and users can also utilize the indirect information provided by others, such as ratings and tone. Previous studies have found that a combination of both indirect and direct reputation evaluation can improve overall predicted reputation accuracy [30]. The goal of our reputation model is to assist decision-making by using past behaviour and indirect information as a predictor of likely future behaviour.

Since it is hard to obtain reliable personal information while recruiting participants in the real world (especially in online recruitment scenarios), we attempted to construct a reputation system that didn't rely on the personal background information of participants. Additionally, a rational use of indirect information is to select the source information and weigh it based on the credibility of the provider [31]. We gather the source of indirect information from participants; hence it is necessary to have objective information to support the credibility of the participants. To effectively utilise citizen scientists in the analysis of large volumes of acoustic sensor data, baseline skill data must be gathered and interpreted to assess the performance of individual participants. To do this we calculate an initial reputation score which is made up of both direct and indirect reputation data sources. The reputation model is illustrated in Figure 1. The model involves direct and indirect measures of reputation.

- *Direct Reputation (DR)*: These sources come from participants' past performance. Such reputation information should be regarded as the most trustworthy support, because it does not involve any subjective concerns and may not be masqueraded.

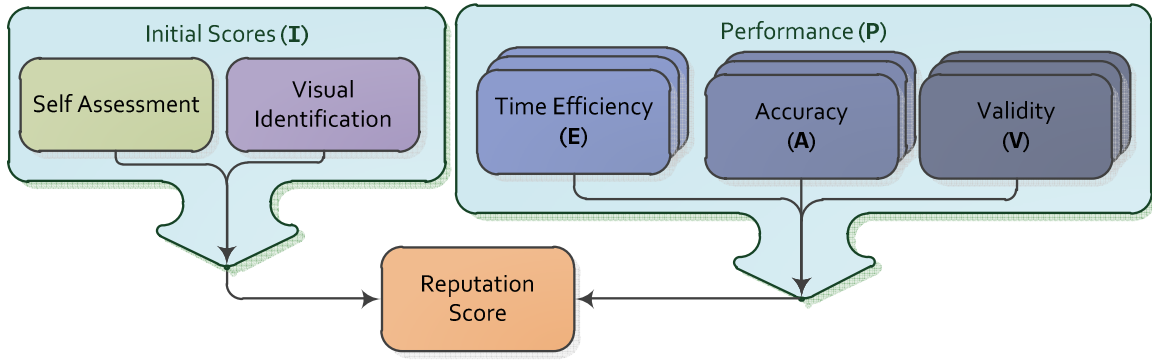


Figure 1 - Proposed Reputation Model

- *Indirect Reputation (IR)*: Indirect reputation, or so called annotation acquisition, is generated by all participants. Hsueh, Melville, and Sindhwan [32] comment that annotation acquisition is able to be of great assistance to supervised information management. This kind of reputation information includes all subjective thinking about the targeted participant.

To obtain baseline scores (**I**) for each participant, participants were required to complete a self-assessment questionnaire and species visual identification quiz. The tests rated the participants existing identification skills subjectively and also evaluated their ability to visually identify some common species. Answers from the tests were converted into weighted scores by using a weighted average. Due to the subjective nature of the self-assessment, these scores cannot be considered as trusted. Therefore, to further assess credibility an objective test is required – thus the inclusion of the visual identification tests. Moreover, the identification test only required participants to identify specific species chosen by the experiment designers, meaning the test results do not represent the whole perceived identification skills of the participant. Based on these concerns, the weighting assigned to both the identification test and the self-assessment were made equal. From this the initial reputation score of each participant can then be generated. The performance attributes used in our reputation model consisted of: accuracy (**A**) defined as correct answers divided by total species; validity (**V**) defined as correct answers divided by number of answers; efficiency (**E**) defined as recording time divided by time spent annotating. The reputation scores were calculated by a weighted average between **I**, **A**, **V**, and **E**. The exact weighting chosen for each input were calibrated after conducting the experiment.

$$I = w_1 S + w_2 M \quad (1)$$

where $w_1 + w_2 = 1$

$$P_n = w_3 A_n + w_4 V_n + w_5 E_n \quad (2)$$

where $w_3 + w_4 + w_5 = 1$
and n is the n^{th} experiment

$$R_n = w_6 I + w_7 \left(\frac{1}{n} \sum_{0 < k \leq n} P_k \right) \quad (3)$$

where $w_6 + w_7 = 1$

The reputation (**S**) scores were calculated by a weighted average of **I**, **A**, **V** and **E**. However, to begin with we did not have the weights for calculating the reputation score. We set the performance results for participants conducting audio annotation as **P** and reputation score as **S**.

We hypothesise that there will be a positive correlation between Initial Score and the participant's trial performance.

B. Productivity Tools

Components of the experiment were included to demonstrate that the basic efficiency of annotators can be improved through a more effective design of the tools they use to assist in species identification. This idea aims to explore simple ways of sorting and filtering data that although naïve will still provide value to participants by improving their ability to accurately identify species in acoustic recordings.

Premise

A call reference library is available to annotators while they are identifying bird species using our online acoustic workbench [8]. This library is comprised of common vocalisation of over 200 bird species common to South Eastern Queensland. Species vocalisations were identified and selected using the experience of trained ornithologists and from existing literature [33]. The library provides users with a tool to compare vocalisations in recordings (visually and aurally) with a set of common calls identified by experts. The reference library is very similar to a digital field guide, albeit filled with examples of audio from the real world.

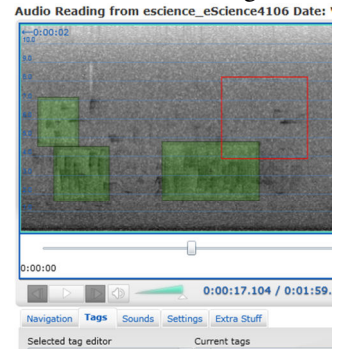


Figure 2 - The MQUTeR Sensors Tagging Tool

The problem with this tool however is there is no way to optimise searching of the library when trying to identify a species. The library does have

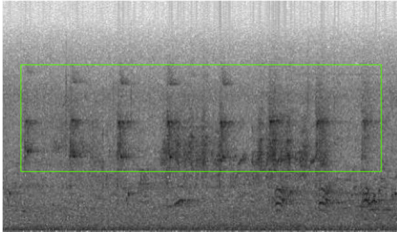
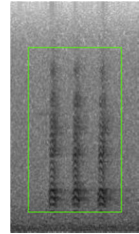
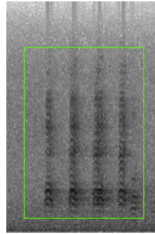
AKP – Australian King Parrot	
	
TC1 – Torresian Crow (3 cries)	TC2 – Torresian Crow (4 cries)
	

TABLE 1- EXAMPLE TAGS

some basic sorting features and a name search to aid this process; however searching for an unknown species by name is fundamentally flawed - as you cannot search for a name you don't know.

Instead, in order to make the annotation tool more useful a novel way of searching the library was devised. Using the information already available as part of the annotation process (i.e. frequency bounds and duration in time), a simple n-dimensional Euclidean distance classifier was devised. The classifier made use of three metrics defined by the annotator as they tag a section of a spectrogram.

Discovering a species vocalisation within a spectrogram and tagging that vocalisation is defined here as annotation. The tag links some arbitrary meta-data (usually a species name) with the time and frequency bounds of the acoustic event (the vocalisation). Figure 2 shows some tags – rectangular regions defined by their bounds. These bounds include a lower frequency (bottom), an upper frequency (top), and duration (width). Since a tag can occur in any point in time, using its start time (left edge) as a bound is not appropriate. Thus duration was used to represent both start and end points of a tag. However the frequency bounds of a tag are limited to a strict domain - thus they can be used independently.

Tool design

The Euclidean distance calculation is a simple way to measure how far two points are away from each other in Cartesian space. Multiple dimensions can be used to represent each variable being compared. The formula used is [34]:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (4)$$

Where $\mathbf{p} = (p_1, p_2, \dots, p_n)$ and $\mathbf{q} = (q_1, q_2, \dots, q_n)$ are two points in Euclidean space and n is the number of dimensions those points occupy.

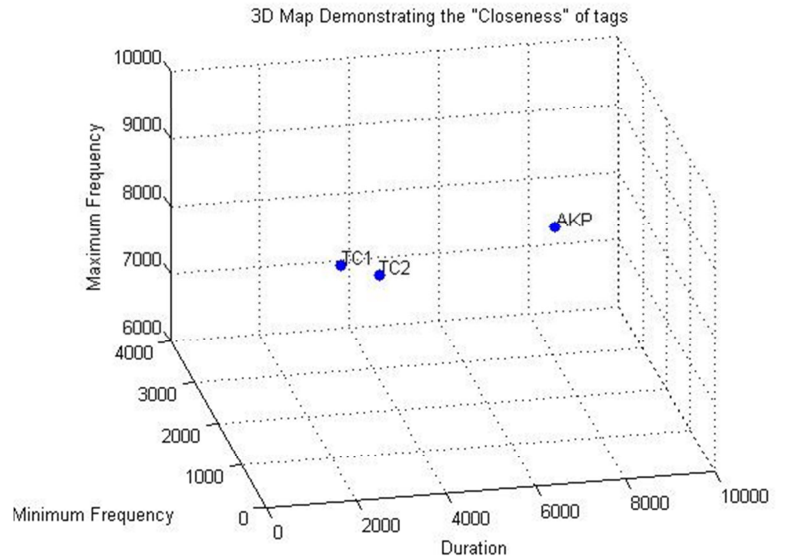


Figure 3 – A 3D map of example tags

A practical example can be seen in TABLE 1 and Figure 3. Presented are three tags - bounded annotations of a spectrogram. Two of these tags (TC1) and (TC2) look very similar, whilst the third, (AKP), looks very different. By measuring how far the bounds of these tags are away from each other in Euclidean space (Figure 3) it is possible to rank how similar they are.

When applied to our annotation tools we believe that sorting tags stored in the reference library on “how far” they are apart (from the sound they are trying to identify) will allow our participants annotate audio data more efficiently and more accurately.

This tool was named “FESB” – Find Events with Similar Bounds.

IV. EXPERIMENT

A. Process Design

15 participants were selected with perceived avian species identification skills ranging from expert to novice. It is important to note that this experiment was purposely scoped as a pilot study: an effort to determine if our theories had merit. We conducted the experiment with as many people as we could with our allotted resources.

An initial trust level needed to be established for participants. Initial trust levels were based on three metrics: (a) a subjective perceived skill level, (b) a visual species identification test to obtain an objective measure of the participant’s ability to detect species from visual stimuli, and (c) an objective measure of a participant’s ability to annotate audio data. The result was used to determine the initial reputation scores of participants and to predict performance in the subsequent audio annotation experiments. Audio annotation experiments are designed to utilise a participant for analysis of acoustic sensor data in order to identify the vocalisations of the species within.

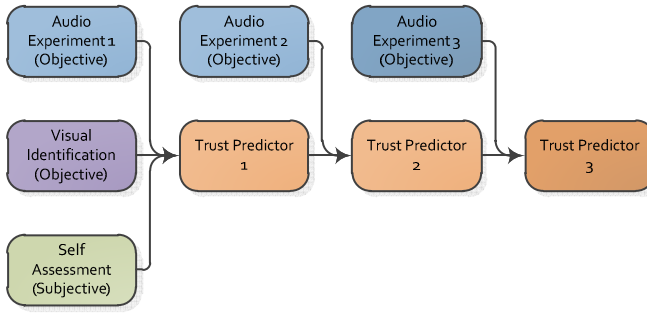


Figure 4 - Basic Experiment Process

Figure 4 displays a simple diagram of the steps involved:

- Step 1 - complete the self-assessment,
- Step 2 - complete the species identification quiz,
- Step 3 - analyse 3 x 2 minute segments of acoustic sensor data and identify each unique species.

The acoustic sensor data component of the experiment involved subjects analysing three, two minute segments of acoustic sensor data. Each segment contained approximately the same number of unique species (15 – 20) however species composition varied between each 2 minute segment. The time taken to complete the analysis for each segment was recorded and identification accuracy determined upon completion of the analysis. For the first and second 2 minute segment, participants were asked to analyse the segments by annotating each call heard and seen on a spectrogram with the name of the species. A call reference library was made available to the subjects to assist in identification. For the final 2 minute segment, participants were asked to analyse the recording by annotating each call, utilising a semi-automated tool (the FESB tool) to assist in identifying calls, along with reference call library.

More details on the trust reputation system and the comparison tool implementation follow.

B. Initial Predictions

Initial predictions of performance and identification accuracy were determined by results from the self-assessment task, species identification quiz, and an initial annotation task. Based on initial assessment, participants were initially classified into the following skill levels: Beginner, Intermediate, and Expert.

The participants are grouped this way so that the experiment can highlight different trends in these important groups. For example, we expect Experts to consistently perform efficiently and we expect to see improvement in the Beginners group's performance – especially when they are allowed the use of the FESB tool.

Interestingly before the trial was run the participants were more evenly distributed across skill levels. After including the trial with the self-assessment and identification quiz we noticed that general performance was reduced significantly. Initial participant classifications are presented in TABLE 2.

We surmise this change in distribution occurred because the audio annotation task is comparatively difficult.

TABLE 2 - INITIAL PARTICIPANT CLASSIFICATION

Skill Level	Beginner	Intermediate	Expert
Cut offs	0.00% \geq $x > 33.33\%$	33.33% \geq x $> 66.66\%$	66.66% \geq x $> 100.00\%$
Pre quiz trust	7	4	4
Trial trust	11	4	0

C. Calculating Weights

To establish if a correlation existed between Initial Score (**I**) and Trial Performance (**P₁**) the results of each experiment were graphed and a linear regression analysis calculated. The visual inspection of the result

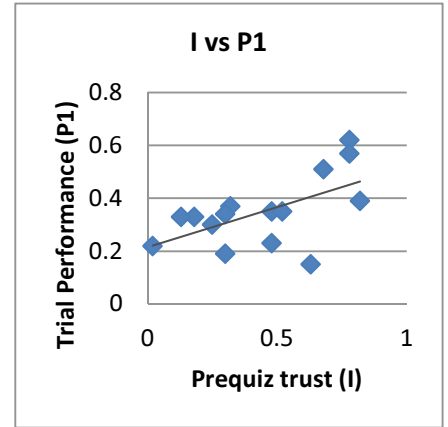


Figure 5 - Initial trust score vs. first Performance

demonstrated a rough positive correlation of the data (Figure 5).

Following this, the weights for both variables (otherwise known as predictors) were calculated. The weights are required to measure how well the initial reputation (**R₁**) predicts the participant's performance in experiment two (**P₂**).

To do this we use regression analysis to determine the relationship between the input variables. This process set the weights between **I** and **P₁** as equal to begin with. The regression calculation determined the coefficients between the variables that would fit the data well. Coefficients and an R-Squared value (indicating how good the fit was to the data) were returned. The coefficients calculated in this table map to weightings used in the reputation formula (5). TABLE 3 shows the numbers returned from regression:

TABLE 3 - COEFFICIENTS FOR I AND P AND THEIR FITTED R-SQUARED VALUE

Coefficient 1 – w_6	0.1453
Coefficient 2 – w_7	0.8547
R-Squared value	0.74

The higher an R-squared the better the coefficients can calculate results that emulate the input data. The R-Squared value for our reputation model indicated an acceptable fit to the experimental data.

Getting the best weighting for all dimensions is meaningful because we expect the reputation score to be representative of the actual performance result. Thus, we input the weightings for the initial score and the performance trial into the reputation equation (see equation (3)) resulting in equation (5).

$$R_1 = w_6 I + w_7 \left(\frac{1}{n} \left(\sum_{0 < k \leq n} P_k \right) \right) \quad (5)$$

$$= 0.1453 * I + 0.8547 * (P_1)$$

By comparing the initial reputation (R_1) with the performance in experiment two (P_2) it is possible to predict performance with an average of 90% accuracy (see TABLE 4). This is a slight (2.5%) improvement over just using performance alone to predict the next performance of a participant.

The standard deviation for all variables is reasonably low. Thus we place a high confidence on the results calculated.

TABLE 4 – GENERAL REPUTATION PREDICTING PERFORMANCE

	I	P ₁ (Trial)	R ₁	P ₂
Mean	0.45	0.35	0.36	0.40
StDev	0.26	0.13	0.14	0.13
P₁ vs. P₂ Correctness		R₁ vs. P₂ Correctness		
87.5%		90%		

From the result shown in TABLE 5, we found that we predicted the performance (average) of beginners less than their actual performance but on average we predicted the performance of intermediate group very closely. However, there is naturally some variance associated with these calculations.

TABLE 5 - CORRECTNESS BY SKILL LEVEL

	Avg. Correctness Difference	Standard Deviation
Beginner	0.07	0.05
Intermediate	0.00	0.07

D. FESB Tool Evaluation

To make the Euclidean classifier tool easy to use a button was created in the existing tools (Figure 2, or see <http://sensor.mquter.qut.edu.au/>). When clicked the Find Event with Similar Bounds (FESB) button copies the information of the tag currently selected by the user, into the reference library's search field and orders the reference library tags by how close they are to the search parameters.

To evaluate the effect of the filtering tool on performance of participants, annotation accuracy and time taken to annotate each three minute audio segment was recorded.

E. Performance Difference

For the efficiency with tooling section of this paper, the statistics calculated are derived several ways. Generally though, an average is taken from the first two experiments (which were identical) and compared to the difference of the third. In this way the first two experiments set a baseline that the third can be compared with.

Following is a table (TABLE 6) that breaks the performance of each group down for the first two experiments.

TABLE 6 - AVERAGE PERFORMANCE OF PARTICIPANTS FROM EXPERIMENT 1 AND 2

Classification	Avg. Accuracy	Avg. Correct vs. Incorrect	Avg. Time
Beginner	14%	78%	23%
Intermediate	43%	92%	23%
All	27%	85%	23%

TABLE 7 - DIFFERENCES IN PARTICIPANTS' PERFORMANCE BETWEEN EXPERIMENT 3 AND PREVIOUS EXPERIMENTS

Classification	Beginner	Intermediate	All
Avg. Accuracy Improvement	14%	11%	13%
Avg. Correct vs. Incorrect Improvement	4%	0%	2%
Avg. Time Improvement	-1%	2%	1%
Avg. FESB Button Use	10.75	7.86	9.4

Analysis of these results suggests a significant difference in volume of tags input on average. To determine significance in the surge of tags annotated in the third experiment when compared to the previous experiments, a T-test is conducted. The T-test reveals significance with $P > 0.975$ for all users. This means that for beginner and intermediate participants that there is a high probability that the data collected was not incidental. Beginners have the strongest t-test probability at $P > 0.999$; however intermediates did not have such a strong guarantee with their T-test score for accuracy, at only $P > 0.8$.

The results for the other two performance metrics (efficiency and validity) revealed no significant results when analysed statistically. Their means remained constant throughout the experiments with similar variances.

However special cases do exist. Validity for example has an extremely close average and variance in the pre and post-tool experiments, suggesting this value is both accurate and did not correlate with the use of the FESB tool.

Also, efficiency received its largest variance in experiment 3. The data reveals that some participants were quicker by up to 30% whilst others were slower by up to 40% (compared to average past performance). Despite this the average time taken in all experiments did not experience much variance.

V. DISCUSSION

A. Reputation Management

Many citizen science projects recruit participants online and such interaction generates many benefits. However, it also poses a challenge for the development of trust. In a virtual environment, it is difficult to verify individual identities and

their actions cannot be easily sanctioned. Therefore, some reputation systems only consider targeted performance and do not make use of the initial information that can be gathered for a reputation baseline while recruiting. However, our experimental results demonstrate that the consideration of initial information is meaningful and beneficial. In fact on average we predict a participant's performance with 90% accuracy. This increases our confidence in the analysis the participants conduct. The more reputation information we hold the more confidence we will be able to place in our participants.

B. Tooling

The aim of the tooling analysis was to discern any measureable difference in the performance of participants when given better tools. The experiment measures performance in three ways: Efficiency (time taken over length of audio), Accuracy (true positives over true positives and true negatives), and Correctness (true positives over false negatives and true positives).

This study predicted an improvement in participant performance when provided with an additional tool designed to make audio annotation work easier (the FESB button). However, no discernable difference in the time taken to analyse or the rate of correct answers vs. wrong were detected. In fact both of these measures of performance had no statistically significant changes.

What was observed however was a substantial increase in the number of tags participants tagged during a session when using the FESB tool. This result was backed up by a T-Test with a strong probability indicator ($P > 0.975$). When evaluated this means on average participants took the same time, had the same ratio of correct vs. wrong, but tagged more annotations. In fact every participant in the experiment had a positive change in the number of annotations completed when using the FESB tool. The effect observed was even stronger in beginners where on average, beginners tagged 14% more tags than their previous experimental averages.

Many participants undertook the experiment whilst being monitored by the paper's authors. We noticed that in general, experiment three was met with enthusiasm from participants after being shown how the FESB tool worked (part of the standard experimental protocol). In the first two annotation activities, many participants felt frustrated due to the difficult nature of audio annotation. In fact, some participants even quit – upset with the fact that audio annotation requires a steep learning curve. Thus when presented with a tool that ordered the reference tags library for them – even when explained as only a simple and rough tool – the participants all demonstrated a renewed desire to give acoustic annotation more attention.

The exception to the general behaviour was observed in the experts that participated in the experiment. The two most highly skilled participants (the experts) did not use the FESB button at all. As the experts were well versed in avian acoustics, they told us they had no need for the tool. The data collected seemed to back up this statement – their performance scores were very consistent.

Participatory Sensing and Participatory analysis are concepts that rely on getting participants to devote their time to a project; these participants may not have formal qualifications and are usually not experts. As such the challenge of any system in this research area is to help the participants who do not yet have the necessary skills to reliably participate. Despite not seeing any improvement in efficiency or correctness, we did see a measurable increase in volume of annotations. We associated this increase with participants perceiving they had a chance to operate within a complex domain when provided with better tools.

C. Suggested Experiment Improvements

While analysing the experiment data a significant amount of variability was detected. We suspect this variability can be accounted for by implementing the changes detailed below. We theorise that because the suggested improvements below were not implemented in our experiment, that it may explain why the initial experiment predictions were not observed.

We propose that a dedicated testing and experiment platform is necessary. We found that the instructions issued were often not read or followed accurately by our participants. We suggest participants (especially beginners) found it hard to follow the instructions since the annotation task was generally hard for them.

Providing a dedicated platform solves issue's that arose from participants timing their annotation session (one less task), and confusing navigation issues – which arise from being instructed to use part of an entire website. Additionally allowing a progressive save of experiment completion will allow the experiment to be more flexible and easier to complete in segments of smaller work.

This paper conducted a relatively small study with few participants. We propose conducting the experiment with many more participants will reduce variability and provide better results. There is a lot of different information that can be collected from this experimental layout which we did not get the chance to use due to variability.

VI. FUTURE WORK

The prediction of reputation baseline for new participants and dedicated work into designing appropriate tools for annotation are both problems that will greatly benefit future work.

Our proposed reputation metric incorporates direct and indirect sources of information and aggregates them by assigning weights that express their importance. As a future work we would like to investigate the viability and application of this metric in real world citizen science projects. Investigating different types of information which can generate subjective and objective support is also a necessary future task; the more information that is made use of, the more support the reputation score can provide.

Tools for annotating faunal acoustic data need improvements so participants' can perform at higher levels of efficiency when annotating. We believe that many opportunities exist to improve tools for these participants. We particularly see opportunity in utilising data currently ignored

(but still available) by many data collection projects. More comparative experiments need to be conducted with more variables and more participants. In brief, we need more intelligent tools. We will also look into the cognitive effects that affect a participant's behaviour when annotating.

In our opinion, the work done in this paper is worth investigating in a similar, larger scale, experiment. We think better results will be observed by increasing sample size and improving the experiment design.

VII. ACKNOWLEDGMENTS

The Microsoft QUT eResearch Centre is funded by the Queensland State Government under a Smart State Innovation Fund (National and International Research Alliances Program), Microsoft Research and QUT. This research was conducted with the support of the QUT Institute of Sustainable Resources and the QUT Samford Ecological Research Facility.

VIII. REFERENCES

- [1] IUCN, "IUCN Red List version 2010. Numbers of threatened species by major groups of organisms (1996 - 2010). http://www.iucnredlist.org/documents/summarystatistics/2010_IR_L_Stats_Table_1.pdf. (accessed on 23 February 2011).", ed, 2010.
- [2] A. Underwood, "On beyond BACI: sampling designs that might reliably detect environmental disturbances," *Ecological applications*, vol. 4, pp. 3-15, 1994.
- [3] S. H. Gage, *et al.*, "Assessment of ecosystem biodiversity by acoustic diversity indices," *The Journal of the Acoustical Society of America*, vol. 109, pp. 2430-2430, 2001.
- [4] T. Penman, *et al.*, "A cost-benefit analysis of automated call recorders," *Applied Herpetology*, vol. 2, pp. 389-400, 2005.
- [5] J. Porter, *et al.*, "Wireless Sensor Networks for Ecology," *Bioscience*, vol. 55, pp. 561-572, 2005/07/01 2005.
- [6] M. A. Acevedo and L. J. Villanueva-Rivera, "Using Automated Digital Recording Systems as Effective Tools for the Monitoring of Birds and Amphibians," *Wildlife Society Bulletin*, vol. 34, pp. 211-214, 2006.
- [7] T. A. Parker, III, "On the Use of Tape Recorders in Avifaunal Surveys," *The Auk*, vol. 108, pp. 443-444, 1991.
- [8] J. Wimmer, *et al.*, "Scaling Acoustic Data Analysis through Collaboration and Automation," in *2010 IEEE Sixth International Conference on e-Science*, 2010, pp. 308-315.
- [9] M. Depraetere, *et al.*, "Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland," *Ecological Indicators*, vol. In Press, Corrected Proof, 2011.
- [10] M. A. Acevedo, *et al.*, "Automated classification of bird and amphibian calls using machine learning: A comparison of methods," *Ecological Informatics*, vol. 4, pp. 206-214, 2009.
- [11] S. Anderson, *et al.*, "Template-based automatic recognition of birdsong syllables from continuous recordings," *Journal of the Acoustical Society of America*, vol. 100, pp. 1209-1219, 1996.
- [12] R. Bardeli, *et al.*, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognition Letters*, vol. 31, pp. 1524-1534, 2010.
- [13] S. Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," *Bird Conservation International*, vol. 18, pp. S163-S173, M3 - 10.1017/S0959270908000415, 2008.
- [14] J. Cai, *et al.*, "Acoustic sensor networks for environmental monitoring," in *Proceedings of the 5th international conference on Embedded networked sensor systems* Sydney, Australia, 2007, pp. 391-392.
- [15] Z. Chen and R. Maher, "Semi-automatic classification of bird vocalizations using spectral peak tracks," *The Journal of the Acoustical Society of America*, vol. 120, p. 2974, 2006.
- [16] C. Juang and T. Chen, "Birdsong recognition using prediction-based recurrent neural fuzzy networks," *Neurocomputing*, vol. 71, pp. 121-130, 2007.
- [17] C. Kwan, *et al.*, "Bird classification algorithms: theory and experimental results," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04)*. 2004, pp. V-289-92 vol.5.
- [18] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2740-2748, 1997.
- [19] P. Somervuo, *et al.*, "Parametric Representations of Bird Sounds for Automatic Species Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 2252-2263, 2006.
- [20] J. Sueur, *et al.*, "Rapid Acoustic Survey for Biodiversity Appraisal," *PLoS ONE*, vol. 3, p. e4065, 2008.
- [21] J. Burke, *et al.*, "Participatory Sensing," in *ACM Sensys workshop on WorldSensor-Web (WSW'06): Mobile Device Centric Sensor Networks and Applications*, pp. 117-134, 2006.
- [22] J. Goldman, *et al.* (2009, 17/06/2010). Participatory Sensing - A citizen-powered approach to illuminating our world [White paper]. [White paper]. Available: http://wilsoncenter.org/topics/docs/participatory_sensing.pdf
- [23] Galaxy Zoo. (2010, 7/7/2010). *The Story So Far*. Available: <http://www.galaxyzoo.org/story>. <http://www.galaxyzoo.org/team>
- [24] C. B. Cooper, *et al.*, "Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy," *Bioscience*, vol. 59, pp. 977-984, 2009.
- [25] K. L. Huang, *et al.*, "Preserving privacy in participatory sensing systems," *Computer Communications*, vol. 33, pp. 1266-1280, 2010.
- [26] S. Reddy, *et al.*, "Evaluating participation and performance in participatory sensing," *UrbanSense08, November*, vol. 4, 2008.
- [27] A. Abdulmonem and J. Hunter, "Enhancing the Quality and Trust of Citizen Science Data," in *IEEE Sixth International Conference on e-Science*, 2010, pp. 81-88.
- [28] M. W. Towsey and B. Planitz, "Title," unpublished.
- [29] X. Wang and F. Zhang, "A New Trust Model Based on Social Characteristic and Reputation Mechanism for the Semantic Web," in *Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on*, 2008, pp. 414-417.
- [30] D. Houser and J. Wooders, "Reputation in Auctions: Theory, and Evidence from eBay," *Journal of Economics & Management Strategy*, vol. 15, pp. 353-369, 2006.
- [31] S. Ruohomaa, *et al.*, "Reputation management survey," in *Second International Conference on Availability, Reliability and Security*, 2007, pp. 103-111.
- [32] P.-Y. Hsueh, *et al.*, "Data quality from crowdsourcing: a study of annotation selection criteria," presented at the Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, Boulder, Colorado, 2009.
- [33] A. Keast, "Song Structures and Characteristics: Members of a Eucalypt Forest Bird Community Compared," *Emu*, vol. 93, pp. 259-268, 1993.
- [34] StatSoft. (2011, 03/05/2011). *Electronic Statistics Textbook*. Available: <http://www.statsoft.com/textbook/cluster-analysis/>