This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

Sonification to Support the Monitoring Tasks of Security Operations Centres

Louise Axon, Jassim Happa, Alastair Janse van Rensburg, Michael Goldsmith, and Sadie Creese

***Abstract*—Sonification (the representation of data as sound) may offer a solution to some of the network-security monitoring challenges faced in security operations centres (SOCs). Prior work has shown that sonification can present network-security information to humans effectively, and indicated that security practitioners foresee potential for sonification to aid in scenarios related to their work. The use of sonification by security practitioners in tasks relevant to SOCs has not been examined, however. To address this gap, we assessed the use of sonification by security practitioners in network-security monitoring tasks in an experimental setting. We report on the results of a study in which we compared the performance of security practitioners using a Security Information and Event Management (SIEM) tool with their performance using a SIEM tool that incorporated sonification, in a primary and a non-primary monitoring task. In both tasks, a number of aspects of the monitoring performance of participants were significantly improved when sonification was used. Our results support the potential for sonification to aid in SOC tasks, and indicate a need to validate the utility of sonification systems by running them in operational SOCs.**

I. Introduction

As organisations continue to be targeted by cyber-attacks increasing in volume, variety and sophistication, monitoring network security is crucial to preventing the organisational harm that could result from breaches. The focus of this article is on the security-monitoring work carried out in security operations centres (SOCs). SOCs aim to mitigate cybersecurity threats towards the organisations for which they are responsible [21]. Security practitioners working in SOCs interact with a range of automated tools, such as intrusion-detection systems (IDSs), to detect and mitigate malicious network activity. Security Information and Event Management (SIEM) tools are widely used to present the information produced by these automated detection tools to security practitioners. The provision of effective tools to support security practitioners is essential, given the huge volumes of complex, multivariate network data to be monitored, and the pressure of the job resulting from the constant threat of malicious activity.

In this article, we explore the potential for using sonification (the representation of data as non-speech audio [14]) alongside existing monitoring tools (in particular, the SIEM tool) to aid in the network-security monitoring tasks of practitioners working in SOCs. Sonification has been experimented with in a number of fields, and has been shown to offer benefits to some monitoring tasks carried out by humans. A key advantage of sound as a medium for presenting information is that humans are able to monitor sound peripherally, meaning that the monitoring does not require their direct focus. It was shown that business-process monitoring could be carried more effectively as a non-primary task using a continuous sonification, than using a visual presentation, without affecting performance in a separate primary task, for example [15].

Some prior research has indicated the ability of humans to use sonification to understand information pertaining to network security, and to "hear" network-attack conditions [4], [7]. It was shown that humans could detect certain network-attack conditions (including denial-of-service attacks and botnet activity) more accurately using a sonification system than an IDS [10] and than three leading anomaly-detection systems that use machine learning [11]. The user testing reported on this topic has been limited, however, and none has targeted our intended users (SOC practitioners), or focused on the specific network-security monitoring tasks and setups of SOCs.

A number of contexts in which sonification could have the potential to aid in SOC working practice have been identified, through interviews on the benefits sonification could offer in theory in light of the challenges faced in SOCs [2]. The use of sonification by security practitioners in these SOC tasks has not yet been explored: only hypothetical cases have been made for the potential utility of the approach. Our aim in the study reported in this article was to explore the use of sonification by security practitioners, and to assess whether using sonification alongside a SIEM tool in a scoped set of these contexts improved the network-security monitoring performance of security practitioners in an experimental setting.

In this article, we present a SIEM tool incorporating sonified network-packet and IDS-alerting information. We provide an assessment of the extent to which the sonified information improved the performance of security practitioners monitoring as both a primary and a non-primary task, in an experimental setting. Our contribution is evidence of the improvement sonification may offer to monitoring performance in such tasks, qualified by the views of security practitioners on the utility of such systems and the challenges that might arise in integrating them into their work in SOCs.

II. Study Background and Aims

Security practitioners working in SOCs include security analysts (whose responsibilities can include preliminary detection, triage of events, and responding to customer tickets) and security engineers (who are responsible for maintaining infrastructure and creating detection rules, for example). SOCs often operate for 24 hours a day, with practitioners required

The authors are with the Department of Computer Science, University of Oxford, Parks Road, Oxford, OX1 3PR (email: {*firstname.lastname*}@cs.ox.ac.uk).

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

to work long shifts, including night shifts, looking at multiple screens for extended time periods (often with multiple monitors, positioned on desks and around the SOC [20]). While automated detection and mitigation tools are widely used in SOCs, security practitioners play a vital role in interpreting the network-security state, identifying false-positive alerts and also threats that may have been missed by automated tools, and making mitigation decisions based on the criticality of an apparent attack and the risk appetite of the organisation [9].

The pressure and demanding nature of SOC work have been highlighted in Human-Computer Interaction (HCI) research [19], [20]. Pressure arises from the constant possibility of malicious activity. Timeliness in the detection of such activity is of paramount importance to maintaining organisational security. Threats to the network must be recognised quickly, and dealt with in a way matching the needs of the organisation, to minimise harm. Ongoing research into the provision of approaches to aid in SOC work include the development of automated techniques (using machine learning, for example), and the identification of solutions relating to human factors.

Axon et al. presented contexts of use and design requirements for sonification in SOCs [2]. Our aim in the study we report on in this article was to explore the use of a sonification system incorporating some of these design developments, in two of these contexts. Firstly, network-security monitoring in real time as a **primary task**, in which monitoring is the sole focus of the practitioner's attention. Secondly, network-security monitoring in real time as a **non-primary task**, whilst the practitioner conducts other work-related tasks. The latter context was proposed based on the benefit that the potential to monitor sounds relating to network security peripherally could offer to multitasking practitioners in busy SOCs.

As well as the sonification of network packets, we aimed to assess the use of sonified alerts, which may be useful to security practitioners when monitoring in the targeted contexts [2]. We therefore developed a sonification system that represented network-packet information and information about IDS alerts simultaneously, which is presented in Section III.

We aimed to assess the use of sonification alongside tools that are used to carry out network-security monitoring tasks in real-world SOCs. This was important in order to ensure that the study made as realistic an assessment as possible of the utility of sonification as an addition to the tools that are actually deployed in SOCs. From existing literature on SOC working practice, and previous discussions with security practitioners, we observed that most security practitioners working in SOCs carried out their network-security monitoring work using a SIEM tool [2], [20]. We therefore focused on assessing the use of sonification alongside a SIEM tool.

Our aim was to explore the utility of the SIEM tool that incorporated sonification compared with a SIEM tool with no sonification included (the **S-SIEM** and the **SIEM**, respectively, as presented in Section III) in the two contexts of use: monitoring in real time as a **primary task**, and as a **non-primary task**. We aimed to assess the monitoring capability of participants in each of these contexts, using the two tools. We considered the monitoring capability of participants to be their ability to accurately and efficiently detect (recognise the presence of), and accurately identify (understand the nature of) packet-traffic anomalies (anomalous deviations in the network-packet traffic, which may be related to network attacks), and alerts (generated by Snort IDS).

We present the detailed hypotheses, and assess them using the results of the study, in Section VI, avoiding repetition in order to save space in this article and improve readability. As well as assessing monitoring capability in these ways, we aimed to explore the views of participants on the study, the suitability of sonification systems for carrying out tasks in the SOC environment, and the potential for integrating sonification into SOCs. We summarise our findings in Section VII.

III. Development of Tools for the Study

We designed two tools: the **SIEM** replicated commercial SIEM dashboards; while the **S-SIEM** was identical to the **SIEM**, but included sonification. We elected to produce our own tools for the purpose of this study, rather than using an existing commercial SIEM, for example, in order to a) avoid the bias that might result from participants' familiarity with a particular SIEM (we aimed to create a SIEM that captured properties observed across a range of commercial SIEMs), and b) enable us to incorporate elements relating to sonification (which we judged would be more practicable in a SIEM of our own creation than in a commercial SIEM).

Whilst in real-world use, SIEM tools would usually also incorporate other information, such as threat-intelligence sharing information and correlations across monitoring tools, the representation of packet traffic and IDS alerts were the focus of this study. We therefore restricted the SIEM and the sonification to represent these two types of information, in order that the study was not complicated by extraneous information. We aimed to produce a **SIEM** and **S-SIEM** that were identical apart from the additional elements relating to sonification. This meant that we could explore the effect of using sonification alongside a SIEM tool without changing other aspects of it, so that any effects observed could reasonably be attributed to the use of the elements relating to sonification, rather than being due to some other difference between the two tools.

A. *SIEM Tool*

In the design of the **SIEM**, we aimed to replicate the visual representations of data (both as text, and as plots or visualizations) observed across existing commercial SIEM tools. We drew on a number of existing commercial SIEMS which we know to be market-leading. We examined the types of visualization present in sample dashboards produced for those SIEMs [8], in order to develop a SIEM as close to those that might be used in SOCs as possible. We also considered the displays available in packet-capture tools such as Wireshark.

In Table I we show the elements we selected for inclusion in the SIEM dashboard, and the existing commercial SIEM displays that informed our selection of each element. Some of the plots were directly replicated from the commercial SIEM displays, while others were designed differently, with the aim of displaying the same type of information.

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

TABLE I: Selection of SIEM dashboard elements based on commercial tools

Dashboard element		**Inclusion in commercial tools**
Protocol counts plot:	a time series of the traffic on the network, divided by protocol/application	NetIQ, Splunk
Incoming protocols plot:	a time series of the traffic incoming to the network (sent from an IP address external to the network to an internal network IP address), divided by protocol/application	IBM
Outgoing protocols plot:	a time series of the traffic outgoing from the network (sent from an internal network IP address to an IP address external to the network), divided by protocol/application	IBM
Alert counts plot:	a time series of the alerts generated, coloured according to alert severity	McAfee, IBM, Splunk
Packets table:	information about the packets captured on the network (packet time, source and destination IP addresses and ports, protocol or application, and packet size), presented in a table	Wireshark
Alerts table:	information about the alerts generated (alert time, class, severity, associated source and destination IP address, and protocol or application), presented in a table	LogRhythm, NetIQ, McAfee, ArcSight, Splunk
IP parallel coordinates:	plot showing connections (coloured according to protocol or application) between source and destination IP addresses	McAfee
Port parallel coordinates:	plot showing connections (coloured according to protocol or application) between source and destination ports	McAfee

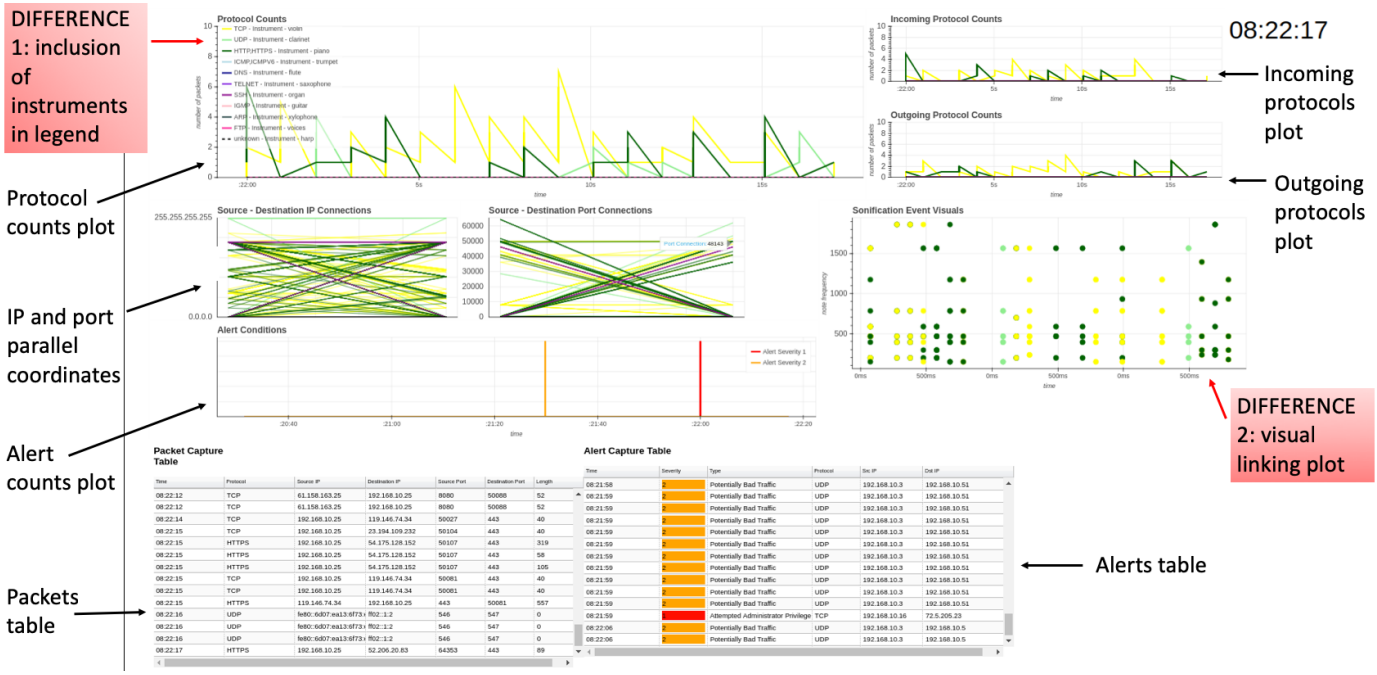


Fig. 1: **S-SIEM** dashboard created for the study

Guidelines for the display of network-security information supported our decision to use line plots and bar charts to represent the time series described in Table I [6]. We aimed to adhere to visualization design principles in the design of the SIEM dashboard. Many design principles relate to the design of interactive visualizations; we drew on only those that could be applied to non-interactive visualizations, since the videos we used in the study were not interactive. An important principle was to create consistency across the visual representations, in order to enable users to more easily become familiarised with a standard way of displaying information [17]. We aimed to be consistent across the plots and tables in our use of colours for representing certain information (different protocols and applications, and different severities of alerts, for example). We also aimed to be consistent in the placement and format of legends, labelling of plots, and formatting of information such as timestamps.

B. *S-SIEM Tool*

We aimed to replicate the **SIEM**, incorporating sonified information and visual elements relating to the sonified display.

1) Sonification approach: In designing the sonified display, we used the model presented by Axon et al. for the sonification of network-security data [5]. In accordance with that model, we defined a tuple describing the sonification of packets and

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

alerts combined, in terms of the *data channels* to be represented (packets and alerts), and the continuous and discrete *dimensions* of these channels (e.g., protocol as a dimension of packets, and severity as a dimension of alerts). This tuple enabled us to create a space of the possible mappings from relevant network-security data parameters to sound.

Our selection of mappings from data to sound, and of the scaling functions with which they would be implemented, was informed by previous Magnitude Estimation studies into the effectiveness and design of data-sound mappings [3], [22]. Based on these prior results, we selected a mapping from packet rate to tempo (scaled with power function 0.53), packet size to articulation (scaled with power function 0.24), and alert severity to pitch. Alerts with severity 1 and 2 were mapped to two different musical motifs, where the pitch of the motif representing alerts with severity 1 was higher than that representing alerts with severity 2. We deliberately kept the sonified alert information involved in this study simple, to avoid presenting participants with too much information to feasibly be used in the study, particularly since packet-traffic information was also being sonified.

We fed these data-sound mappings into the data-sound mappings space, and thus selected the following *relations*, describing the sonification of network-security data.

- *Data channels*:
  - Packet → individual note
  - Alert → musical motif
- *Data dimensions (continuous)*:
  - Rate (packet) → tempo (note)
  - Size (packet) → articulation (note)
  - Severity (alert) → pitch (motif)
- *Data dimensions (discrete)*:
  - Type of information (packet) → pitch range (note)
  - Common-ness (packet) → consonance (note)
  - Direction (packet) → spatialisation (note)
  - Protocol (packet) → instrument (note)

"Type of information" refers to the aspect of the packet about which information is being represented (e.g., IP address or port), while "common-ness" refers to how common its observed value is to the network. For IP addresses, for example, we used "hotlists" of the most commonly observed IPs on the network over a baseline period, then assessed whether each IP address observed in real time was present in this hotlist and therefore "common". Common values were mapped to consonant notes (that formed part of the harmonic progressions of the music), while uncommon values were mapped to dissonant notes (that were not within these harmonic progressions).

2) Incorporating sonification into the SIEM: While we aimed to replicate the **SIEM** dashboard as closely as possible in the **S-SIEM**, there were two adjustments we made to the dashboard relating to the addition of sonification. Firstly, we added information linking instruments to the protocol or application they represented to the *protocol counts plot* legend, in order to enable users to make a link between the instruments heard and the protocols represented (by colour) using the protocol-instrument mapping. Secondly, we added a plot to aid users in linking the sounds heard to their meaning in terms of the data (this has previously been cited as a potential requirement for sonification tools that could be useful to SOCs [2]). We ensured that the placement of all other dashboard plots was identical in the **SIEM** and **S-SIEM**, and the presence of this extra plot was the only difference.

We created the **S-SIEM** dashboard presented in Figure 1, on which we have indicated (by highlighting in red) the two ways in which it differed visually from the **SIEM**. In the **S-SIEM**, as data appeared in the SIEM dashboard, it was represented sonically in the sonified display.

IV. Methodology

We present the methodology we devised to address the aims presented in Section II. We describe the study preparation and process, and our methods for data collection and analysis.

A. Study Preparation

We created network-attack datasets and set up Snort IDS for use in the study. We recorded videos of the tools running on these network-attack datasets, and also created videos describing the two tools, for use in training the participants.

1) Network-attack datasets: We created network-attack datasets by extracting sections of the CIC-IDS 2017 Intrusion Detection Evaluation Dataset, which is publicly available online [1]. The dataset contains benign traffic and up-to-date common attacks, synthetically generated and captured on a network testbed over a one-week period [18]. We selected this dataset for the study because of its realisticness (to achieve realistic background traffic, its creators profiled the behaviour of humans interacting with machines) and ease-of-use (the dataset is clearly labelled and documented).

Three network-attack datasets were extracted from the full CIC-IDS 2017 dataset. The *Baseline Dataset* was taken from the "baseline day" for the CIC-IDS 2017 dataset (Monday 9am-5pm). This was a day during which no attacks were synthetically launched on the network, and as such no known attacks were recorded in the dataset. We used the *Baseline Dataset* in the training stage, in order to present participants with "baseline data" running in the two tools.

Two further datasets were used for the study tasks. *Dataset 1* was taken from a time in CIC-IDS 2017 that did not contain any synthetically generated attacks (and was included to assess the ability of participants to recognise a network state with no attacks). *Dataset 2* was taken from the time at which an FTP brute-force attack began in CIC-IDS 2017.

2) Setting up Snort IDS, and alerting on the dataset: We ran Snort IDS version 3 on the network-attack datasets, to generate IDS alerts. In Snort 3, signatures can be written to produce alerts based on network-traffic features. The alerts produced by Snort are usually given a class (describing the type of alert: "Attempted Privelege Escalation", for example), a severity, and include details such as the source and destination IP addresses and ports, and protocols or applications used by the traffic that caused the alert to be generated.

The signatures used were those included in the Snort 3 Community Rules, which we opted to use in order that our installation of Snort was consistent with installations that

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

might be used by other Snort users. Whilst many leading SOCs will also craft their own IDS signatures based on knowledge of the networks they monitor and the threats to them, this is done on a basis individual to SOCs, and replicating this practice for the purpose of this study would have been challenging.

3) Baselining and running the tools: As described in Section III-B, we required a baseline "hotlist" of the most commonly observed IP addresses and ports during "normal" network operation. We therefore calculated hotlists of the source and destination IP addresses and ports present in the *Baseline Dataset*. IP address and port common-ness was calculated according to these hotlists in the real-time sonification of the network-attack datasets for the study.

We used `tcpreplay` to play the network-attack datasets onto a loopback network. On this loopback network, we ran a "Packet Logger" to recapture a sample of the packets (at most 30 per second) and log them in a PCAP file. Packets with properties that were not commonly observed on the network (those using uncommon IP addresses and ports, or using protocols or applications that, unlike UDP, TCP and HTTP, for example, were not commonly used) were sampled with a higher probability to increase their chances of being represented. At the same time, we ran an "Alert Logger" (using Snort IDS) on the loopback network, to capture and log the alerts generated in response to the replayed packets. The logged packets and alerts were then sent directly to the **SIEM** for representation visually and as text. The **S-SIEM** also controlled the sonification of network data and alerts.

4) Creating training and task videos: Prior to running the study, we recorded videos of both the **SIEM** and the **S-SIEM** running while each network-attack dataset (*Baseline Dataset*, *Dataset 1*, and *Dataset 2*) was replayed on the network. We used Kazam screen recorder to capture the screen and corresponding audio. We elected to use pre-recorded videos in the study, rather than running the tools live, in order to ensure that the information presented was consistent across all participants. Since packets were sampled randomly, with a certain probability, by the "Packet Logger" on each run, running the study live might have led to differences in the sampling across participants, resulting in differences in the information displayed in the **SIEM** and **S-SIEM**.

We also created two training videos: a **SIEM** training video and an **S-SIEM** training video. These two videos were used in the training stage of the study, to introduce each tool and describe its functionality to participants. We chose to use videos in order to ensure consistency of the training information: we presented each participant with the same two videos. The videos showed a static screen capture of each tool, and the plots and tables present in each tool were described verbally in the video by the researcher, with the relevant parts of the visuals indicated by computer mouse movements captured in the video. In the **S-SIEM** training video, the sonification approach was described, in addition to the visuals.

B. Study Process

We recruited a convenience sample of 22 participants for the study. Participants were security practitioners with experience of working in SOCs. 4 SOCs participated, both internal SOCs who are responsible for the security of the single organisation they are placed within) and multitenanted (SOCs who monitor security on behalf of multiple client organisations). The study took place in-person with each participant at their organisation, in a room separate to the SOC. We received ethical approval for this study from the Central University Research Ethics Committee at the University of Oxford.

In the materials we used in the recruitment process we deliberately avoided mentioning sonification, or the fact that we were aiming to compare approaches newly developed by the researchers. Our aim here was to avoid biasing participants prior to the study: the knowledge that we were assessing the use of tools that we had designed ourselves might have caused acquiescence bias, in which the participant aims to provide the results they believe the researchers are hoping for.

1) Introduction: We began by introducing the premise of the study to each participant. Participants then read an information sheet, which described the study aims, the processes that would be followed in the study, the measures the researchers would take to ensure that data collected were treated ethically, and the procedures to be followed should a participant have any concerns during or after completing the study. Participants indicated their consent to participate and to be audio recorded by signing a consent form. We took audio recordings from this point onwards, in order to capture the dialogue. Participants then answered preliminary questions on paper, which captured demographics including the job role and network-security monitoring experience of the participant, the types of network-security monitoring tools (including SIEM tools) they were accustomed to using, and their level of musical experience.

Participants sat at a desk, next to the researcher. A Macbook Pro laptop and a monitor were set up on the desk, and headphones (Bose SoundLink Around-Ear headphones) were used by the participant to listen during the training videos and the study tasks involving sonification. We used headphones because of the restrictions of visiting organisations: we could not play sound out loud in the rooms in the organisations in which we ran the study, since we needed to be unobtrusive to the work of the rest of the organisation. In order to ensure that the sound quality was consistent and did not impact on the results, we took two measures. Firstly, we presented the sounds to all participants through the same headphones during training and study tasks. Secondly, we kept the volume level of both the laptop and the headphones consistent across participants and tasks; therefore the volume of sound cues and alerts, as well as the noise level, were the same for all participants.

2) Training: Participants watched the training videos, first for the **SIEM** and then for the **S-SIEM**. This order of watching was required, since the **S-SIEM** was designed as an extension of the **SIEM**, and it was therefore necessary to explain how the **SIEM** functioned first.

Before starting to watch the videos, participants were informed that they could pause the video at any time to ask the researcher questions. We elected not to use training datasets, in which the participant could practice completing the study tasks, in the interest of keeping the study to a reasonable length of time. Furthermore, we did not wish to train participants in

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

the detection of any particular types of attacks, which would have had to have been present in the training datasets. This enabled us to obtain results on the capabilities of participants using the tools for the first time, and to gather their views on the viability of learning to use these tools.

3) Study task overview: Participants were divided into four groups, using one of the two tools to monitor one of the two datasets in each study task.

A Participants in *Group A* completed *Study Task 1* using the **SIEM** running on *Dataset 1*, and *Study Task 2* using the **S-SIEM** running on *Dataset 2*.

B Participants in *Group B* completed *Study Task 1* using the **SIEM** running on *Dataset 2*, and *Study Task 2* using the **S-SIEM** running on *Dataset 1*.

C Participants in *Group C* completed *Study Task 1* using the **S-SIEM** running on *Dataset 1*, and *Study Task 2* using the **SIEM** running on *Dataset 2*.

D Participants in *Group D* completed *Study Task 1* using the **S-SIEM** running on *Dataset 2*, and *Study Task 2* using the **SIEM** running on *Dataset 1*.

Each participant was assigned to one of these four groups pseudorandomly. We ensured that at least 5 participants were assigned to each group, enabling us to compare performance across groups in our analysis. We aimed to avoid bias through this pseudorandomisation. For example, if participants always monitored with the **SIEM** prior to monitoring with the **S-SIEM**, it might be expected that their performance would improve over time as a result of becoming accustomed to the task, which might bias the results to suggest that participants had performed better using the **S-SIEM**.

At the beginning of each study task, participants were shown the *Baseline Dataset* run through either the **SIEM** or **S-SIEM** – whichever had been selected for the participant in that task. As in the rest of the study, we used pre-recorded videos to present the tools running on the *Baseline Dataset* at this stage. The aim was to enable participants to experience each tool running in real time, and to observe the types of traffic that were present in a network "baseline".

During the **SIEM** baseline video, the researcher described verbally aspects of the presentation described in the training videos: the balance of protocols in the plots, and the presence of new alerts in the tables, for example. The **S-SIEM** baseline video was paused, to allow the researcher to highlight these aspects (since the headphones would have obstructed this communication while the video was playing).

Participants then completed the study task, listening using headphones when the **S-SIEM** was used. Videos were played from a MacBook Pro laptop, presented in a larger monitor such that all aspects presented were easily visible.

4) Study Task 1: monitoring as a primary task: Participants were presented verbally with the following instructions:

*You are required to monitor network security as a primary task using the (**SIEM/ S-SIEM**). Please monitor using this tool, and respond verbally if you observe either anomalous packet traffic, or a severity 1 or 2 alert, as follows:*

- *If you observe anomalous packet traffic, say: "anomalous traffic", and give a verbal description of the anomalous traffic you have observed.*
- *If you observe a severity 1 alert, say: "alert severity 1".*
- *If you observe a severity 2 alert, say: "alert severity 2".*

To aid participants, we provided them with a sheet of paper on which the required verbal responses were written: "*anomalous traffic*", "*alert severity 1*", and "*alert severity 2*".

Participants then completed the task, monitoring the video played from the Macbook Pro and shown in the larger monitor. At the beginning of each task, the researcher spoke the word "go" into the audio recording, in which participants' verbal responses were collected. This enabled us to later calculate the time at which participants made each verbal response in relation to the dataset being represented (and thus to calculate, for example, the time delay between the presence of a severity 1 alert and the response of the participant having identified it).

5) Study Task 2: monitoring as a non-primary task: Participants were presented verbally with the same instructions as in *Study Task 1*, with the following information added: "*Simultaneously, you are required to carry out a separate primary task. This task involves the creation of folders and files in the OSX terminal*".

For the separate primary task, participants were presented with an instruction sheet, containing instructions such as "*create a new folder named 'new folder 1'*" and the corresponding text to be typed in the terminal (e.g. "`mkdir "new folder 1"`"). This instruction sheet was presented in the MacBook screen, with the OSX terminal open alongside it ready for the participant to carry out the primary task. We assumed that participants would be familiar with the types of terminal operations required in the primary task, and this proved to be the case in the study – we checked with participants whether they were comfortable with this type of task before beginning *Study Task 2*. The instruction sheet included the exact commands needed, in case these were not known. Before beginning the study task, participants were given the opportunity to practice the primary task, which some participants took.

This separate primary task was selected to be representative of the type of task that a practitioner might realistically be required to carry out whilst monitoring the network. For the purposes of this study, this separate primary task was appropriate insofar as we judged it realistic, and it did not create any ethical concerns (as writing an email or an incident report might give away more information on the participant's personal details or working practice, for example).

The non-primary monitoring task was then presented as a video played from the MacBook Pro (in the terminal of which participants completed the separate primary task), and presented in the same larger monitor as in the previous task. The exact multitasking requirements were not specified to participants: the amount of attention to devote to the primary and the secondary task was left to be decided by the participant. The aim here was to capture as realistic as possible a view of the way in which practitioners might multitask in such a scenario, and the amount of attention they might devote to each task (although this realisticness was undoubtedly limited by the constraints of the study: the unfamiliarity of the tools and setup to participants, compared to in their day-to-day work).

6) Post-study interviews: After the study tasks had finished, participants completed a spoken post-study interview on topics

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

including their experience in the study, the realisticness of the study design, and their thoughts on using sonification in their SOC work. The interview was semi-structured: the researchers had a prepared set of questions as a guide, and selected questions to ask based on the direction in which the conversation had been taken during the study itself (if a participant had given their view on how realistic the study setup was earlier in the study, for example, we did not repeat a question on this topic in the post-study interview).

We also recorded and analysed the spoken dialogue between the participant and researcher throughout the study. In this way, in addition to the views expressed in the post-study interview, we were able to evaluate the views that participants expressed throughout the study. This included comments on the tools presented, the study tasks, the utility of sonification and other factors relating to its effectiveness (such as sonification design), and the realisticness of the monitoring setup.

C. Analysis

We aimed to evaluate the effect of incorporating sonification into a SIEM tool on primary- and non-primary task monitoring performance. This was achieved by comparing the monitoring performance of participants using the **SIEM** and that of those using the **S-SIEM** during *Study Task 1* and *Study Task 2*.

While we used quantitative analysis to compare the performance of participants using the two tools in *Study Task 1*, and in *Study Task 2*, separately, we did not use quantitative analysis to compare performance across the two tasks. This was because, since *Study Task 1* was always carried out before *Study Task 2*, we could not account for the possible effect of each participant first carrying out *Study Task 1* (using one of the two tools) on their performance in *Study Task 2* (using the other of the two tools). The quantitative comparisons of **SIEM** and **S-SIEM** in *Study Task 1*, and in *Study Task 2*, comprised comparisons of a number of assessments of monitoring performance, which are described in Table II.

We transcribed the interview data collected throughout the study and in the post-study interview, and analysed it using template analysis (a qualitative analysis technique in which the researcher pre-forms ideas of the themes to be examined, and adds to these themes as new evidence arises) [16].

1) Measuring participants' performance: Participants were asked to respond verbally if they observed an anomalous deviation in the traffic (and give a description of its nature), and if they observed an alert (and describe its severity). Depending on the event type (attack, event, or anomalous deviation), we varied our interpretation of what constituted a detection, as described in Table II. Using these measurements, we assessed detection accuracy and efficiency, and identification accuracy.

2) Comparing detection accuracy and efficiency, and identification accuracy: As shown in Table II, we analysed the accuracy of detection using combinations of precision, recall and F-score [13]. The reasons why each type of assessment was chosen for each event type are explained in the next subsection. We removed outliers in the time taken to detect alerts and the FTP brute-force attack, by removing results for which the z-score had an absolute value greater than 3.29 [12].

We compared the efficiency with which events were detected by the two populations (participants using the **SIEM** and the **S-SIEM**) using t-tests (preceded by an F-test, to decide the type of t-test required based on equality of variance [12]). We compared identification accuracy by calculating the number of correctly and incorrectly identified, and unidentified detections of each type of event. Idenfication rate was calculated as the number of correctly identified true-positive detections, divided by the total number of true-positive detections.

3) Note on approach to network-traffic analysis: attacks and anomalous deviations: As shown in Table II, we used precision, recall and F-score to assess the accuracy of alert detection. Making these calculations requires measurements of the number of true- and false-positive detections and of the number of false-negative (missed) detections. It was straightforward to identify the times at which alerts did and did not occur in each dataset, and we were therefore able to make these measurements with certainty. Our observations during the study on the nature of network-traffic monitoring and on the methods by which this is practiced in SOCs supported a different approach to the analysis of attack detection, in which we analysed attack recall, and explored the network-traffic awareness of participants throughout the datasets.

Participants were asked to monitor for "anomalous traffic", having observed the *Baseline Dataset* represented in the **SIEM** and **S-SIEM**. If, as for the detection of alerts, we had assessed the accuracy of attack detection using precision, recall and F-scores, the responses of participants would have been compared against the ground truth that *Dataset 1* contained no attack, while *Dataset 2* contained an attack after 115 seconds. During the study, however, we observed that participants made anomalous traffic detections extra to those related to the attack.

Furthermore, the comments made by participants indicated that deciding whether anomalous traffic relates to an attack is not necessarily an activity that is immediately carried out whilst monitoring in real time (e.g., "*I think in reality you would look at that and stop/pause, and look again*" (P11)). The decision on whether anomalous traffic is malicious or benign would often be informed by further investigation of the traffic observed, exploring the traffic in closer detail retrospectively. Given that our study explored real-time monitoring using videos, and there was no facility for participants to return to events and investigate them retrospectively, this key decision-making aspect of attack detection was not captured.

Based on these observations, we considered the anomalous-traffic detection capabilities that were important to this study. For the initial real-time part of the monitoring, it is important that the anomalous traffic relating to the attack is noticed, so that security practitioners are made aware that further retrospective investigation of it is required. As such, rather than measuring the accuracy of participants' decisions on whether anomalous traffic relating to a network attack had or had not been observed, we chose to focus on ensuring that participants detected the anomalous traffic at the time of the attack.

We posit that other detections of anomalous traffic, not necessarily relating to an attack, are part of the real-time monitoring process. Such a detection might be returned to retrospectively by practitioners in reality, in order to make

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

TABLE II: Quantitative comparisons made in the study analysis, for network event types: FTP attack; alert; traffic deviations

	Assessment	Details of measurement	Assessment criteria
FTP attack	Detection accuracy (recall) Detection efficiency Identification accuracy	For the FTP brute-force attack (in *Dataset 2*), in the calculation of detection accuracy and efficiency, and identification accuracy, we made the assumption that a participant's first response following the FTP brute-force attack was a detection of traffic or an alert relating to that attack. The first traffic relating to the attack was at 115 seconds, and the first alert at 119 seconds, in the dataset. We counted any response after 115 seconds as a detection of the attack, as all anomalous traffic deviations after that point were ongoing FTP traffic relating to the attack.	**Detection accuracy**: • If the response constitutes a detection of any of the three types of event (detection interpreted for each event type using the assumptions above) → true-positive detection • If the response does not constitute a detection → false-positive detection • If the response is not made within the detection constraints (i.e. the event is missed) → false-negative detection **Detection efficiency**: If true-positive detection: • Time difference between the onset of an event and the first response constituting a detection of it by the participant → detection efficiency **Identification accuracy**: If true-positive detection: • If the attack is correctly described → correct identification • If the attack is incorrectly described → incorrect identification
Alerts	Detection accuracy (precision, recall, F-score) Detection efficiency Identification accuracy	For alerts, we assumed that a participant's response was a detection of the last alert present in the dataset. We observed that some participants took longer to detect alerts, particularly when multitasking, and the participants themselves noted, whilst monitoring, their delay in spotting the alerts ("*Ive also only just noticed that we have sev 2s and sev 1s*" (P5), for example). We wished to capture these late detections in our analysis, so chose not to impose a time limit on responses that would constitute detections of an alert.	
Traffic deviations	Detection accuracy (precision) Detection efficiency Identification accuracy	We assumed that a participant's response was a detection of the last traffic deviation present in our ground truth, if that traffic deviation fell within the 10 seconds prior to the response. Whilst, as described, we chose not to use such a time limit for alert detection, we judged that detection of anomalous traffic was more ambiguous and required some time constraint (given the volumes and variability of the packet traffic, compared with only two variants of alert). As such, treating a response made 20 seconds after the presence of a traffic deviation as a detection of it may not have been appropriate, for example.	

decisions on whether it constitutes an attack. P3, for example, noted that while not necessarily indicating an attack, certain traffic would usually prompt further investigation: "*things like Telnet and FTP would be... not necessarily something that was going to be bad, but something that would at least make me want to look into it further*". Rather than making the same measurements of true- and false-positive and negative detections of anomalous network traffic as we had for alerts, therefore, we judged it more appropriate to assess how the anomalous traffic detections made by participants related to the events occurring in the dataset at the time.

We therefore focused on measuring the following two aspects of monitoring for anomalous network traffic.

a Detection of anomalous traffic at the time of the attack. We calculated the recall accuracy – which accounts for true-positive and false-negative detections, and is a measure of the probability that when the attack occurred, a participant detected it. Recall does not account for false-positive detections, so was unaffected by the detections of anomalous traffic made by participants at other times in the dataset.

b "Network-traffic awareness", by which we describe the relevance of the anomalous deviations in traffic described by participants, to the network traffic present in the dataset at the time. This means, for example, if a participant described an anomalous spike in FTP traffic at a time outside the FTP brute-force attack time window, we would verify whether there was an increased amount of FTP traffic at that time. If the descriptions of anomalous traffic given by participants aligned with the traffic represented in the dataset (in terms of precision, timeliness, and correctness of identification), we considered that this constituted "network-traffic awareness".

The ground truth used in point *a* above was the timing of the attack, which occurred from 115 seconds onwards in *Dataset 2*. We assessed whether, and how efficiently and accurately, participants detected the anomalous traffic and alerts related to the attack. For the assessment of network-traffic awareness (point *b* in the list above), we compiled a list of the ground truth for each dataset. These ground-truth lists were generated iteratively: we began with the anomalous traffic noticed by the researchers: spikes of particular protocols, and times at which particularly dissonant sound was generated, for example.

Because of the large amount of highly varied traffic the datasets contained, we could not guarantee that all anomalous traffic deviations had been captured by the researchers. Therefore, as new anomalous traffic was identified by participants, we verified its presence in the dataset. If the traffic was present at that time, we added it to the ground-truth lists. In cases where anomalous traffic was detected but its nature was not identified, if the traffic did not match the timing of an existing anomalous traffic deviation in the ground truth, we judged that there was too much ambiguity to add it to the ground truth.

We assessed whether correctly detected anomalous traffic deviations were also correctly identified. Responses were

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

coded as either correctly identified (an attempt at identification was made, and matched the ground truth according to our analysis approach), incorrectly identified (an attempt at identification was made, but did not match the ground truth according to our analysis approach), or unidentified (no attempt was made at identifying the nature of the traffic). Identification of the nature of the anomalous traffic is an important part of the network-traffic awareness we assessed. We present the identification rates, and plots of the times at which correctly and incorrectly identified, and unidentified detections were made, in Section V-D.

V. Results

We present the demographics of the participants, and a quantitative analysis of their performance in the study tasks.

A. Participants and Demographics

Of the 22 participants in the study, 20 completed the entire study, while 2 (P4 and P9) were shown the tool-introduction videos and then took part in the interview (due to time constraints of those participants). The demographics of the participants are presented in Table III. All participants had no hearing or visual impairments, and all participants had experience of using a SIEM tool. 13 participants had experience of using security visualizations, while 9 did not.

B. Detection and Identification of the FTP Brute-Force Attack

1) Detection accuracy and efficiency: For each participant, we assessed whether anomalous traffic and alerts had been detected following the occurrence of the FTP brute-force attack at 115 seconds in *Dataset 2*, and if so we calculated the detection time. The results for the first detection of both traffic and alerts (presented separately) relating to the attack by each participant are presented in Table IV.

We analysed the time taken by each participant to make the first detection of either traffic or alerts following the attack. Here, we identified which detection had been made first (anomalous traffic or alert) after the attack had started. The time taken to this first detection by each participant (following the beginning of the attack at 115 seconds in *Dataset 2*), and the way in which the detection was made (whether anomalous traffic or alerts were described) are presented in Table V.

In Table VI, we present the recall rate and detection time (mean and standard deviation) for the detection of anomalous traffic and alerts (separately) following the FTP brute-force attack in *Dataset 2*, using the results presented in Table IV. We also present the recall rate and detection time (mean and standard deviation) for the FTP brute-force attack using both traffic and alerts. Here, we use the results presented in Table V, considering the first detection of either traffic or alerts (whichever was first) following the beginning of the attack.

We ran t-tests to check for significant differences between participants using the **SIEM** and the **S-SIEM** in the mean time taken to detect either traffic or alerts relating to the FTP brute-force attack (whichever the participant detected first, as presented in the "Overall" row of Table VI).

For *Study Task 1*, we selected a two-sample t-test assuming equal variances (since an initial F-test showed that there was no significant difference between the variances of the two populations). The results of the t-test showed that there was no significant difference in the mean times taken by participants in *Study Task 1* using the **SIEM** and the **S-SIEM** to their first detection of either traffic or alerts (whichever was first) following the attack, with significance $\alpha = 0.05$, since *t Stat* $= -0.72$ and *t Critical two-tail* $= 2.31$, and therefore *t Stat* $>$ *-t Critical two-tail* and *t Stat* $<$ *t Critical two-tail*.

For *Study Task 2*, we selected a two-sample t-test assuming unequal variances (since an initial F-test showed that there was a significant difference between the variances of the two populations). We found that the mean time taken by participants in *Study Task 2* using the **S-SIEM** to detect traffic or alerts (whichever was first) following the attack was significantly faster than that taken by participants using the **SIEM**, with significance $\alpha = 0.05$, since *t Stat* $= 3.37$ and *t Critical two-tail* $= 2.31$, therefore *t Stat* $>$ *t Critical two-tail*.

2) Identification accuracy: No participant identified that the attack was an FTP brute-force. It is likely that further information and investigation would have been needed for this specific attack identification to be made, and this was supported by the comments of some participants (P1 commented on the extra information needed to identify the attack as an FTP brute-force, for example). A number of participants did, however, identify a continuously high level of FTP traffic, which was a key indicator of the attack. In Table VII, we show the proportion of the participants who detected anomalous traffic following the attack, who also made this identification.

C. Detection and Identification of Snort IDS Alerts

1) Detection accuracy and identification accuracy: In Table VIII, we present results on accuracy of IDS-alert detection and identification. The table reports the precision, recall and F-score measures of detection accuracy, and the identification rate (the proportion of the alerts detected that were also correctly identified). As shown, participants obtained perfect recall of alerts using the **S-SIEM**, meaning there were no missed (false-negative) detections. Using the **SIEM**, the recall rate shows that there were some false-negative detections, resulting in lower (but still reasonable) F-scores of 0.95 in *Dataset 1*, and 0.97 in *Dataset 2*. The mean identification rate across both tasks was slightly higher for participants using the **SIEM** (0.99) than the **S-SIEM** (0.96).

2) Detection efficiency: In Tables IX and X we present the time (mean and standard deviation) taken to true-positive alert detections by participants using each tool in the two study tasks. In Table IX these results are divided over *Dataset 1* and *Dataset 2*, while in X the results are calculated from the detection times across both datasets.

We ran t-tests to check for significant differences between participants using the **SIEM** and the **S-SIEM** in the mean time taken to detect alerts in both *Study Task 1* and *Study Task 2*. For *Study Task 1*, we selected a two-sample t-test assuming unequal variances. The mean time taken by participants using the **S-SIEM** to detect alerts in *Study Task 1* was significantly

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

TABLE III: Participants' demographics: gender, age bracket, job role, years of job experience, and level of musical experience

Gender		**Age**		**Job Role**		**Years**		**Musical Experience**	
Male	19	18-24	3	Senior Security Analyst	6	< 1	1	No musical training, does not actively listen to music	2
Female	3	25-34	10	Level 1 Security Analyst	3	1	4	No musical training, does actively listen to music	12
		35-44	7	Level 2 Security Analyst	1	2	5	Some instructed musical training	2
		45-54	2	Level 3 Security Analyst	1	3	3	Some self-taught musical training	3
				Security Analyst (level not given)	4	4	2	Graded standard in a musical instrument	3
				SOC Manager	2	5	2	Advanced (e.g., conservatoire-level) musical training	0
				Network Security Engineer	2	6	1		
				Threat Intelligence Analyst	3	7	0		
						8	0		
						9	1		
						10+	3		

TABLE IV: Detection of, and time taken (in seconds, *s*) to detect, anomalous traffic and alerts (presented separately) following the FTP brute-force attack by each participant (P)

	Traffic				Alerts			
	SIEM		**Son. SIEM**		**SIEM**		**Son. SIEM**	
	Detected	*s*	Detected	*s*	Detected	*s*	Detected	*s*
Task 1	✓(P3)	5	✓(P2)	4	✓(P3)	9	✓(P2)	4
	✓(P10)	4	✓(P7)	11	✓(P10)	20	✓(P7)	3
	✓(P15)	30	✓(P14)	5	✓(P15)	3	✓(P14)	2
	✓(P17)	0	✗(P18)		✓(P17)	2	✓(P18)	5
	✓(P22)	3	✗(P20)		✓(P22)	13	✓(P20)	2
Task 2	✓(P5)	8	✓(P1)	0	✓(P8)	4	✓(P1)	2
	✓(P16)	19	✓(P11)	1	✓(P12)	6	✓(P6)	0
	✓(P21)	7	✓(P13)	4	✓(P16)	5	✓(P11)	5
	✗(P8)		✓(P19)	5	✓(P21)	2	✓(P13)	9
	✗(P12)		✗(P6)		✗(P5)		✓(P19)	7

TABLE V: Detection and time taken (in seconds) to detection following the FTP brute-force attack: first detection by each participant (P) – traffic (T) or alerts (A)

	SIEM		**S-SIEM**	
	Detected	Time	Detected	Time
Study Task 1	✓(P3)(T)	4	✓(P2)(T)	4
	✓(P10)(T)	4	✓(P7)(A)	7
	✓(P15)(A)	7	✓(P14)(T)	5
	✓(P17)(T)	0	✓(P18)(A)	5
	✓(P22)(T)	3	✓(P20)(A)	2
Study Task 2	✓(P8)(A)	4	✓(P1)(T)	0
	✓(P12)(A)	6	✓(P6)(A)	0
	✓(P16)(A)	9	✓(P11)(T)	1
	✓(P21)(A)	6	✓(P13)(T)	4
	✓(P5)(T)	8	✓(P19)(T)	5

TABLE VI: Recall rate, and mean and standard deviation (S. D.) of detection times (in seconds) following the FTP brute-force attack: traffic only, alerts only, and overall (first detection of attack-related traffic or alerts)

		SIEM			**S-SIEM**		
		Recall	Detection Time		Recall	Detection Time	
			Mean	S. D.		Mean	S. D.
Traffic	*Task 1*	1.0	3.0	2.16	0.60	6.67	3.79
	Task 2	0.60	7.50	0.71	0.80	2.50	2.38
Alerts	*Task 1*	1.0	6.75	5.19	1.0	3.20	1.30
	Task 2	0.80	4.25	1.71	1.0	4.60	3.65
Overall	*Task 1*	1.0	3.60	2.51	1.0	4.60	1.82
	Task 2	1.0	6.60	1.95	1.0	2.0	2.35

faster than that taken by participants using the **SIEM**, with significance $\alpha = 0.05$, since *t Stat* $= 3.06$ and *t Critical two-tail* $= 2.05$, and therefore *t Stat* $>$ *t Critical two-tail*.

For *Study Task 2*, we selected a two-sample t-test assuming unequal variances. The mean time taken by participants using the **S-SIEM** to detect alerts in *Study Task 2* was significantly faster than that taken by participants using the **SIEM**, with significance $\alpha = 0.05$, since *t Stat* $= 3.59$ and *t Critical two-tail* $= 2.10$, and therefore *t Stat* $>$ *t Critical two-tail*.

TABLE VII: Identification of the FTP brute-force attack: proportion of those participants who detected anomalous traffic following the attack, who also correctly identified the traffic as continuously high levels of FTP

	SIEM	**S-SIEM**
Task 1	3/5	0/3
Task 2	1/3	3/4

TABLE VIII: Precision, recall, F-score and identification accuracy for alert detections in the study tasks

		SIEM			**S-SIEM**		
		Dataset 1	*Dataset 2*	Mean	*Dataset 1*	*Dataset 2*	Mean
Task 1	Precision	1.0	1.0	1.0	1.0	1.0	1.0
	Recall	0.8	1.0	0.9	1.0	1.0	1.0
	F-score	0.89	1.0	0.95	1.0	1.0	1.0
	Identification	1.0	1.0	1.0	1.0	0.93	0.97
Task 2	Precision	1.0	1.0	1.0	1.0	1.0	1.0
	Recall	1.0	0.88	0.94	1.0	1.0	1.0
	F-score	1.0	0.94	0.97	1.0	1.0	1.0
	Identification	1.0	0.93	0.97	1.0	0.90	0.95
Both	Precision	1.0	1.0	1.0	1.0	1.0	1.0
	Recall	0.9	0.94	0.92	1.0	1.0	1.0
	F-score	0.95	0.97	0.96	1.0	1.0	1.0
	Identification	1.0	0.97	0.99	1.0	0.91	0.96

TABLE IX: Time taken to true-positive alert detections (mean and standard deviation) in each dataset

	SIEM				**S-SIEM**			
	Dataset 1		*Dataset 2*		*Dataset 1*		*Dataset 2*	
	Mean	S. D	Mean	S. D.	Mean	S. D.	Mean	S. D
Task 1	4.60	2.88	3.29	3.09	1.60	1.14	1.75	1.73
Task 2	7.33	3.88	5.25	3.96	3.14	2.34	2.28	2.22
Both	6.17	3.43	8.30	7.07	2.50	2.02	2.05	1.99

TABLE X: Time taken to true-positive alert detections (mean and standard deviation) across both datasets

	SIEM		**S-SIEM**	
	Mean	S. D	Mean	S. D.
Task 1	3.63	2.32	1.78	1.64
Task 2	6.20	3.78	2.44	2.24

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

D. Anomalous Traffic Deviations: Network-Traffic Awareness

We present an analysis of participants' detection and identification of anomalous deviations in the traffic throughout both datasets (rather than only at the time of the attack in *Dataset 2*, which we presented the analysis of in Section V-B). For this study, we considered an anomalous traffic deviation to be a change in the traffic to a state different from the most frequently present "baseline". This baseline traffic was, for the most part, a balance of IPV4 traffic sent over TCP, HTTP and UDP, with mostly common (hotlisted) IP addresses and ports used. The decision on whether a change in the traffic constituted an anomalous traffic deviation was made by the researcher based on their knowledge of this usual traffic state.

As we described in Section IV-C, we took an iterative approach to creating the ground-truth lists for anomalous traffic against which the responses of participants were assessed. The ground-truth lists we produced through this process, which are the basis on which we analysed the performance of participants in this section, included the following types of traffic deviation: FTP and SSH traffic; increases in TCP; dissonant notes (related to the presence of unusual IP addresses or ports); IPV6 addresses; and traffic sent between high ports.

TABLE XI: Precision of anomalous traffic detection, and identification accuracy

Task	Assessment	**SIEM**			**S-SIEM**		
		Dataset 1	*Dataset 2*	Mean	*Dataset 1*	*Dataset 2*	Mean
Task 1	Detection: precision	0.93	1.0	0.97	1.0	1.0	1.0
	Correctly identified	0.93	0.83	0.88	0.38	0.43	0.41
	Incorrectly identified	0.07	0.06	0.07	0.0	0.0	0.0
	Unidentified	0.0	0.11	0.06	0.62	0.57	0.60
Task 2	Detection: precision	1.0	1.0	1.0	0.73	1.0	0.86
	Correctly identified	0.25	0.71	0.48	0.81	0.96	0.89
	Incorrectly identified	0.0	0.0	0.0	0.13	0.0	0.07
	Unidentified	0.75	0.29	0.52	0.06	0.04	0.05
Both	Detection: precision	0.95	1.0	0.98	0.83	0.94	0.89
	Correctly identified	0.78	0.81	0.80	0.62	0.73	0.68
	Incorrectly identified	0.06	0.05	0.06	0.07	0.0	0.04
	Unidentified	0.17	0.14	0.16	0.31	0.27	0.29

TABLE XII: Time taken to detection of traffic anomalies: mean and standard deviation

Task	**SIEM**				**S-SIEM**			
	Dataset 1		*Dataset 2*		*Dataset 1*		*Dataset 2*	
	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.
Task 1	5.25	3.20	3.76	1.82	2.25	1.14	3.25	2.18
Task 2	6.33	3.51	4.5	2.18	4.21	2.26	1.64	1.26

Table XI shows the precision with which participants detected anomalous deviations in the traffic, and the accuracy with which they identified them. Table XII shows the time taken by participants to detect anomalous traffic deviations following their occurrence in the dataset.

E. The Effect of Musical Experience on Monitoring Performance

We divided participants into two groups: those with and without musical experience. We considered participants with musical experience to be those who reported having some level of instructed or self-taught musical training, and may have reached a graded standard in a musical instrument.

We ran t-tests to check for significant differences between participants with and without musical experience using the **S-SIEM** in the mean time taken to detect alerts, and also in the mean time to detect either traffic or alerts following the FTP brute-force attack (the first detection of either following the attack, as used in Section V-B). For these assessments, we considered only the datasets in which participants had used the **S-SIEM**. We collected all alert-detection times of participants using the **S-SIEM**, and considered FTP brute-force attack detection only for those participants who used the **S-SIEM** in *Dataset 2*, in which this attack was present. We also calculated the accuracy of detection and identification.

1) Network-attack traffic: To assess the effect of musical experience on the time taken to detect traffic or alerts relating to the FTP brute-force attack by participants using the **S-SIEM**, we selected a two-sample t-test assuming unequal variances. The results showed that there was no significant difference in the mean times taken by participants with and without musical experience, using the **S-SIEM**, to their first detection of either traffic or alerts following the attack, with significance $\alpha = 0.05$, since *t Stat* $= -1.57$ and *t Critical two-tail* $= 2.78$, and therefore *t Stat* $>$ *-t Critical two-tail* and *t Stat* $<$ *t Critical two-tail*. Since all participants detected the attack in either the traffic or alerts (as shown in Table V), the recall rates of participants with and without musical experience for the attack were equal (both 1.0).

Of the 3 participants with musical experience who detected the attack traffic (before alerts) using the **S-SIEM**, all 3 correctly identified the attack traffic as continuously high rates of FTP (identification rate 1.0). Of the 3 participants without musical experience who detected the attack traffic (before alerts) using the **S-SIEM**, none correctly identified this attack traffic (identification rate 0). In this study, therefore, participants with musical experience were better able to identify the attack traffic, although the populations compared are too small for the results to be generalisable.

2) Alerts: To assess the effect of musical experience on the time taken to detect alerts, we selected a two-sample t-test assuming equal variances. The results showed that there was no significant difference in the mean times taken by participants with and without musical experience, using the **S-SIEM** to detect alerts, with significance $\alpha = 0.05$, since *t Stat* $= 0.56$ and *t Critical two-tail* $= 2.0$, and therefore *t Stat* $>$ *-t Critical two-tail* and *t Stat* $<$ *t Critical two-tail*. Since, as Table VIII shows, the precision, recall and F-score of participants for the detection of alerts using the **S-SIEM** were all 1.0, it follows that the precision, recall and F-score of participants with and without musical experience for the detection of alerts using the **S-SIEM** were equal (1.0).

Of the 25 true-positive alert detections by participants with musical experience using the **S-SIEM**, 3 were not correctly identified, resulting in an identification rate of $22/25 = 0.88$. 2 of the 44 true-positive alert detections by participants without musical experience using the **S-SIEM** were not correctly identified, resulting in an identification rate of $42/44 = 0.95$.

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

VI. Hypotheses and the Performance of Participants in the Study

We present our study hypotheses, and assess whether each hypothesis was proved true by the results presented in Section V. The assessment of hypotheses is presented in Table XIII, and summarised in Section VI-A. We also compare the results relating to network-traffic awareness across tasks and tools in Section VI-B. We qualify the relevance of all findings by considering the study scope and realisticness in Section VII.

A. Study Hypotheses and Assessment

We summarise the assessment of hypotheses. In each of **Hypotheses A-D**, "*monitoring performance*" is described by three sub-hypotheses, concerning 1) *detection accuracy*, 2) *detection efficiency* and 3) *identification accuracy*. The assessment of these sub-hypotheses is presented in Table XIII.

Hypothesis A: *participants will perform better at monitoring for network-attack traffic as a primary task when using the **S-SIEM** than the **SIEM***. Participants did not perform better when using the **S-SIEM** than the **SIEM** in terms of either detection accuracy, detection efficiency, or identification accuracy (see Table XIII).

Hypothesis B: *participants will perform better at monitoring for alerts as a primary task when using the **S-SIEM** than the **SIEM***. Participants performed better in the detection of alerts (both accuracy and efficiency) when using the **S-SIEM** than the **SIEM**, but using the **S-SIEM** did not improve the accuracy of participants in identifying those alerts detected.

Hypothesis C: *participants will perform better at monitoring for network-attack traffic as a non-primary task when using the **S-SIEM** than the **SIEM***. There was no difference in the accuracy of network-attack detection between participants using the **SIEM** and the **S-SIEM** (participants achieved perfect recall of 1.0 using both tools). Participants detected network-attack traffic significantly more efficiently, and identified it more accurately, when using the **S-SIEM** than the **SIEM**.

Hypothesis D: *participants will perform better at monitoring for alerts as a non-primary task when using the **S-SIEM** than the **SIEM***. Participants performed better in the detection of alerts (both accuracy and efficiency) when using the **S-SIEM** than the **SIEM**, but using the **S-SIEM** did not improve the accuracy of participants in identifying those alerts detected.

Hypothesis E: *a participant's level of musical experience will not affect their monitoring performance in the study*. There was no significant difference in the accuracy and efficiency of participants with and without musical experience in detecting network-attack traffic or alerts using the **S-SIEM**. Participants with musical experience identified network-attack traffic more accurately than those without (although the population sizes in this comparison were small). Contrary to our predictions, participants without musical experience identified alerts slightly more accurately than those with musical experience.

B. Network-traffic awareness: detection and identification of anomalous traffic deviations

We explore the results presented in Section V-D relating to network-traffic awareness. Table XI shows that the mean level of precision in the detection of anomalous network traffic was high for participants using the **SIEM** in both tasks (0.97 in *Study Task 1* and 1.0 in *Study Task 2*), and using the **S-SIEM** in *Study Task 1* (1.0). The high precision shows that when participants stated "anomalous traffic" throughout these tasks, it generally corresponded to an anomalous traffic deviation in the dataset (according to our assumptions, as detailed in Section IV-C). For participants using the **S-SIEM** in *Study Task 2*, whilst multitasking, the level of precision was lower, although still reasonably high (0.86). This may suggest that participants were somewhat misled by the sound, believing that they had heard changes representing anomalous traffic at times when there was no anomalous traffic in the dataset. It is possible that the need to multitask prevented confirmation of the anomalies visually, which might have led to the lower precision using the **S-SIEM** in *Study Task 2* than in *Study Task 1*. This is an area that requires further investigation.

True-positive detections of anomalous traffic deviations were made more efficiently by participants using the **S-SIEM** than the **SIEM** in both *Study Task 1* and *Study Task 2*, as Table XII shows. This suggests that on average participants were more quickly made aware of anomalous traffic deviations when using the **S-SIEM** than when using the **SIEM**, both when multitasking and when monitoring as a primary task.

As shown in Table XI, the highest identification rates for correctly detected anomalous traffic were achieved by participants using the **S-SIEM** in *Study Task 2*, and using the **SIEM** in *Study Task 1* (the proportions of correctly detected traffic that was also correctly identified were 0.89 and 0.88 respectively). The correct identification rates by participants using the **S-SIEM** in *Study Task 1* and the **SIEM** in *Study Task 2* were considerably lower (0.41 and 0.48 respectively). In these cases, a large proportion of the correctly detected traffic was unidentified by participants (no identifying information was included in their response).

Some incorrect identifications were made; amongst the most common was an identification of SSH traffic as Telnet. This may be accounted for by the use of colour in the dashboard: both SSH and Telnet were represented in purple, and while the shades were different, this may have caused some confusion. This incorrect identification suggests an identification performed using the visual display (since it is likely to have been caused by the similarity of these two colours); there were indications of other participants making identifications using the sound itself. P11 and P13, for example, frequently described the instrument and then its meaning in terms of the data when making identifications: "*organ, so therefore SSH, and voices so FTP of some description*" (P13). In future studies, making a quantified assessment of the methods by which users make their identifications of anomalous traffic (using visuals or sound) will be important to understanding the interplay between the use of sonification and visualization.

VII. Discussion

We consolidate our findings in this study. We reflect on the design of the study, identifying limitations and approaches to designing more realistic and effective studies in this space,

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

TABLE XIII: Assessment of Study Hypotheses

Monitoring Aspect and Description	Test	Hypothesis and Assessment
Network-attack traffic		
Detection accuracy: Network-attack traffic will be detected more accurately using the **S-SIEM** than **SIEM**.	$R_{AT}(SIEM) < R_{AT}(Son)$. $R_{AT}(SIEM)$ and $R_{AT}(Son)$ are the recall rates calculated across all k participants in the detection of the attacks at in the set of all attacks AT, using the **SIEM** and the **S-SIEM**, respectively.	**A1**: primary task. **Reject**: $R_{AT}(SIEM) = R_{AT}(Son)$ (both are equal to 1.0, as shown in Table VI). **C1**: non-primary task. **Reject**: $R_{AT}(SIEM) = R_{AT}(Son)$ (both are equal to 1.0, as shown in Table VI)
Detection efficiency: Network-attack traffic will be detected more efficiently (in a shorter time) using the **S-SIEM** than **SIEM**.	$\Delta t_{AT}(SIEM) > \Delta t_{AT}(Son)$. $\Delta t_{AT}(SIEM)$ and $\Delta t_{AT}(Son)$ are the mean averages across all k participants using the **SIEM** and **S-SIEM** (respectively), and across all attacks at in the set of all attacks AT, of the differences $\delta t_{atk}(SIEM) = td_{atk}(SIEM) - tb_{atk}$ and $\delta t_{atk}(Son) = td_{atk}(Son) - tb_{atk}$ (respectively), between the time of the beginning of an attack at, $tb_{atk}(SIEM)$ and $tb_{atk}(Son)$ (respectively), and the time of its detection $td_{atk}(SIEM)$ and $td_{atk}(Son)$ (respectively).	**A2**: primary task. **Reject**: t-test showed no significant difference between $\Delta t_{AT}(SIEM)$ and $\Delta t_{AT}(Son)$. **C2**: non-primary task. **Accept**: $\Delta t_{AT}(SIEM) > \Delta t_{AT}(Son)$ (t-test)
Identification accuracy: Network-attack traffic will be identified more accurately using the **S-SIEM** than **SIEM**.	$I_{AT}(SIEM) < I_{AT}(Son)$. $I_{AT}(SIEM)$ and $I_{AT}(Son)$ are the attack-identification rates calculated across all k participants using the **SIEM** and **S-SIEM**, across all attacks at in the set of all attacks AT.	**A3**: primary task. **Reject**: $I_{AT}(SIEM) > I_{AT}(Son)$ (as shown in Table VII). **C3**: non-primary task. **Accept**: $I_{AT}(SIEM) < I_{AT}(Son)$ (as shown in Table VII)
IDS alerts		
Detection accuracy: Alerts will be detected more accurately using the **S-SIEM** than **SIEM**.	$F_{AL}(SIEM) < F_{AL}(Son)$. $F_{AL}(SIEM)$ and $F_{AL}(Son)$ are the F-scores calculated across all k participants in the detection of the alerts al in the set of all alerts AL, using the **SIEM** and the **S-SIEM**, respectively.	**B1**: primary task. **Accept**: $F_{AL}(SIEM) < F_{AL}(Son)$ (as shown in Table VIII). **D1**: non-primary task. **Accept**: $F_{AL}(SIEM) < F_{AL}(Son)$ (as shown in Table VIII).
Detection efficiency: Alerts will be detected more efficiently (in a shorter time) using the **S-SIEM** than **SIEM**.	$\Delta t_{AL}(SIEM) > \Delta t_{AL}(Son)$. $\Delta t_{AL}(SIEM)$ and $\Delta t_{AL}(Son)$ are the mean averages across all k participants using the **SIEM** and **S-SIEM** (respectively), and across all alerts al in the set of all alerts AL, of the differences $\delta t_{alk}(SIEM) = td_{alk}(SIEM) - tb_{alk}$ and $\delta t_{alk}(Son) = td_{alk}(Son) - tb_{alk}$ (respectively), between the time of the beginning of an alert al, $tb_{alk}(SIEM)$ and $tb_{alk}(Son)$ (respectively), and the time of its detection $td_{alk}(SIEM)$ and $td_{alk}(Son)$ (respectively).	**B2**: primary task. **Accept**: $\Delta t_{AL}(SIEM) > \Delta t_{AL}(Son)$ (t-test). **D2**: non-primary task. **Accept**: $\Delta t_{AL}(SIEM) > \Delta t_{AL}(Son)$ (t-test).
Identification accuracy: Alerts will be identified more accurately using the **S-SIEM** than **SIEM**.	$I_{AL}(SIEM) < I_{AL}(Son)$. $I_{AL}(SIEM)$ and $I_{AL}(Son)$ are the alert-identification rates calculated across all k participants using the **SIEM** and **S-SIEM**, across all alerts al in the set of all alerts AL.	**B3**: primary task. **Reject**: $I_{AL}(SIEM) > I_{AL}(Son)$ (as shown in Table VIII). **D3**: non-primary task. **Reject**: $I_{AL}(SIEM) > I_{AL}(Son)$ (as shown in Table VIII).
The effect of musical experience		
Detection: There will be no significant difference in the ability of participants with and without musical experience to detect attack traffic or alerts.	$a_{AT}(mus) = a_{AT}(nonmus)$, $e_{AT}(mus) = e_{AT}(nonmus)$, $a_{AL}(mus) = a_{AL}(nonmus)$, $e_{AL}(mus) = e_{AL}(nonmus)$. $a_{AT}(mus)$ and $a_{AT}(nonmus)$, and $a_{AL}(mus)$ and $a_{AL}(nonmus)$ are the attack-traffic and alert detection accuracy (respectively). $e_{AT}(mus)$ and $e_{AT}(nonmus)$, and $e_{AL}(mus)$ and $e_{AL}(nonmus)$ the attack-traffic and alert detection efficiency (respectively) of participants with and without musical experience.	**E1a**: network-attack traffic. **Accept**: $a_{AT}(mus) = a_{AT}(nonmus)$ and $e_{AT}(mus) = e_{AT}(nonmus)$ (as shown in Section V-E) **E2a**: alerts. **Accept**: $a_{AT}(mus) = a_{AT}(nonmus)$ and $e_{AT}(mus) = e_{AT}(nonmus)$ (as shown in Section V-E)
Identification: There will be no significant difference in the ability of participants with and without musical experience to identify attack traffic or alerts.	$i_{AT}(mus) = i_{AT}(nonmus)$, and $i_{AL}(mus) = i_{AL}(nonmus)$. $i_{AT}(mus)$ and $i_{AT}(nonmus)$, are the attack-traffic identification accuracy, and $i_{AL}(mus)$ and $i_{AL}(nonmus)$ are the alert identification accuracy, of participants with and without musical experience.	**E1b**: network-attack traffic. **Reject**: $i_{AT}(mus) > i_{AT}(nonmus)$ (as shown in Section V-E) **E2b**: alerts. **Reject**: $i_{AT}(mus) < i_{AT}(nonmus)$ (as shown in Section V-E)

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

in Section VII-A. Section VII-B explores the findings on the integration of sonification into SOCs: its potential utility, practicality, and the design requirements for suitable systems.

A. Study Design: Limitations and Lessons Learnt

1) Study structure and process: There is a potential limitation with the study structure: the comparisons between the primary- and secondary-task monitoring might be biased by the fact that the primary-task monitoring was always carried out first. This could affect performance by enabling participants to become more accustomed to the study processes by the time they participate in the secondary-monitoring task. We judged that presenting the primary-task monitoring before secondary-task monitoring was necessary since the secondary-task monitoring was more complicated (more tasks needed to be completed) and the primary-task monitoring could prepare participants for the secondary-task monitoring to some extent.

The bias should have been minimised by the fact that the same tool was not used for the two tasks by any participant: all participants used one of the two tools in the first study task, and the other tool in the second. Whilst they may have become accustomed to the general processes of the study, therefore, participants would not have become accustomed to using either of the tools, and would not therefore have carried this experience from the primary- to the secondary-task monitoring. Furthermore, we did not quantitatively compare performance across the two tasks, but compared performance using each tool in *Study Task 1*, and then separately compared performance using each tool in *Study Task 2*.

In the analysis of *Study Task 2*, we did not assess the effect of using either tool to monitor, on the performance of participants in the primary shell-based task. Making this assessment of the effect of monitoring on the separate primary task would have required an initial assessment of the performance of each participant completing the primary task only (without monitoring), as in previous similar studies [15]. Our decision not to include this assessment was based on time constraints.

Furthermore, we aimed to capture as closely as possible the actual multitasking methods of the practitioners in their SOC work, and were therefore not prescriptive about how to multitask (how much attention to pay to each display). As some participants noted, in their real SOC monitoring work, if multiple alerts or traffic anomalies were observed they would have stopped trying to complete the separate primary task and focused on the monitoring only. This was reflected by the multitasking approach of some participants in the study, as some did pause their primary-task work at key monitoring points, in order to focus on the dashboard for a time. Given that participants were not given strict instructions on their completion of the primary task, it would not have been meaningful to assess the effect of the monitoring on it. Rather, we explored how having that primary-task distraction affected the monitoring performance, and compared this across tools.

Given the described effects of time constraints on our study design decisions, which may have introduced certain limitations or prevented the exploration of key questions relating to the utility of sonification to SOCs, it is clear that this study must be complemented by longitudinal studies, or studies that involve the use of sonification in operational SOCs.

2) Study setup: We took measures to ensure that the study setup (the tools, network datasets and the monitoring tasks we focused on) was as realistic as possible within practical and time constraints. The views expressed by participants indicated that there were limitations to this, however.

Some aspects of the SIEM were cited as being unrealistic by participants. The SIEM was used for real-time monitoring only in this study and events observed could not be returned to and investigated, in accordance with our study aims to explore real-time monitoring capabilities, and due to the fact that we used pre-recorded videos. Participants noted that their real-time monitoring work using a SIEM would be supplemented by further investigation following the initial recognition of events, and that this was an aspect that was missing from our study. We learned that assessing real-time attack detection is difficult, as this would usually involve investigation after the initial detection of a potential security event.

The observations made by participants on the realisticness of the information included in the SIEM varied. While the inclusion of traffic visualizations was not realistic in some SOCs, who used streams of alerts only as the basis for real-time monitoring, for example, other participants were more accustomed to using these approaches. We designed the SIEM tool based on existing commercial SIEM dashboards. It was suggested, however, that SOCs do not always use the types of visualization-based dashboard produced by SIEM vendors in reality (P18). The types of tools that practitioners are accustomed to using may vary between SOCs, and it is important to take this into account in the design of solutions for SOCs, and of SOC-related studies.

In real-world SOC monitoring tasks, practitioners would have a more in-depth knowledge of the network being monitored than in this study. This may have had an impact on the ability of participants to make decisions on network events. This could be improved in future studies by providing participants with some additional information on the network: a participant (P1) suggested that a diagram of the network architecture would be useful, for example. In including such materials, it is important to avoid information overload. A more natural way of addressing this limitation would be to run studies over a longer time, in which the participants have time to become familiar with the network being monitored, or to run trials of sonification solutions in operational SOCs, on networks with which practitioners are already familiar.

Of the two tasks we focused on this study, monitoring as a non-primary task was considered the more realistic to SOC-monitoring work by a majority of participants. The exception was some Level 1 Security Analysts, who were accustomed to monitoring as a single primary task at times.

3) Measurement and analysis of performance: As described in Table II, we made a number of assumptions in deciding the data event to which a participant's response related. We found that understanding exactly what change in the visuals or sound has led a participant to make a response was difficult, however. For the attack, for example, it was not possible using our study setup to be certain about whether

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

a participant's response was made because they had observed the indicators of that attack condition, or because they believed they had observed something else. We used time windows to reduce the length of data, and therefore number of events, against which a participant's response would be compared, increasing the likelihood that we would assess their response against the event they had actually observed.

A limitation of the time-window approach in constructing the anomalous-traffic ground truth is that events constituting anomalous traffic deviations were frequent (in *Dataset 2* especially), and as such there were stretches of time in which most responses were assessed as relating to some event that had occurred within the last 10 seconds. This meant that the number of false-positive detections was low, and this may have influenced the results by increasing the level of precision calculated. Considering the accuracy of identification is particularly important in this case.

A way of addressing the difficulty of understanding the object of participants' responses may be to capture visual information on the study through video recording, eye tracking, or observation. This would enable physical actions such as pointing at the screen or changes in gaze to be analysed, which may indicate the focus of participants' attention whilst monitoring. Such information could also be valuable in enabling researchers to further understand the methods employed by participants when completing the study. In this particular study, such information could have enabled us to measure how often each participant observed the dashboard whilst multitasking during *Study Task 2*, and thus assess how much attention was devoted to the primary and non-primary tasks, and how participants used the sonification whilst multitasking.

While in this study, for the reasons we presented in Section IV-C3, we elected not to consider the precision of attack detection based on the amount of other anomalous traffic, unrelated to network attack, that was detected, considering the precision of attack detection is of course important. Monitoring tools for SOCs must enable users to make accurate decisions on whether an event observed is malicious and related to an attack, or is benign. Measuring the precision with which sonification can enable this capability is essential to understanding its potential for use as a network-security monitoring approach. Future work should make this assessment, testing the attack-detection precision of users with more substantial training and experience of using sonification, and with the ability to use the tools under study for the type of retrospective exploratory work that is carried out at part of SOC monitoring practice.

B. Using Sonification in SOCs: Utility, Practicalities and Requirements

We found that there were a number of ways in which SOC practitioners, having experienced using the **S-SIEM**, felt that sonification systems could be useful in SOCs. Each context in which sonification had the potential to be useful brought a number of practical challenges.

Of the contexts described by participants, many were connected by their basis in the potential advantage to SOC working practice of using sound to monitor peripherally. Sound is a naturally effective medium through which to present information to be monitored peripherally, given the ability of humans to receive sonic cues without focusing their attention in the direction of the sound source (unlike visual cues, for example). Contexts described that have a basis in this property of sound included using sonification to monitor traffic and alerts as a secondary task, whilst at the desk working, moving around the SOC, or away from the SOC (either on break, or interacting with other departments within the organisation).

Sonification designs for peripheral-monitoring contexts must enable users to interpret the information presented in an accurate and timely way. In Section VI we assessed our study hypotheses, and found that participants monitoring as a non-primary task detected traffic and alerts relating to the network attack more efficiently, and identified it more accurately, when using the **S-SIEM** than when using the **SIEM**. There was no difference in the detection accuracy (assessed using the recall score), since all participants using each tool detected the attack. Furthermore, participants monitoring as a non-primary task detected IDS alerts more accurately and efficiently when using the **S-SIEM** than when using the **SIEM**, but were not more accurate at identifying their severity.

Given that in all but the extra components relevant to the sonification, the two study tools were identical, our hypothesis testing suggests that using sonification may improve the efficiency with which users are able to detect IDS alerts, and traffic and alerts relating to network-attack conditions, when monitoring as a non-primary task. For IDS alerts, recall was higher when participants used the **S-SIEM**, meaning that fewer alerts went undetected. Our assessment of network-traffic awareness also showed that participants more efficiently detected other anomalous deviations in the network traffic when using the **S-SIEM** than the **SIEM**, although their detections were less precise (with a greater number of false-positive detections). The extent to which training with sonification systems can improve the precision of anomalous-traffic detection is a key avenue for future research.

While we have provided evidence supporting the potential benefits of using sonification to practitioners monitoring as a non-primary task, it is important to recognise that we were only able to assess these hypotheses using the results collected in the study. There are number of limitations to making these assessments in an hour-long study: as described in Section VII-A, the time pressure and difficulty of exactly replicating the realistic conditions of SOCs, for example. In light of these limitations, running future longitudinal studies in which sonification systems are actually deployed in SOCs will be vital to understanding their true utility.

It is key that sonification systems can be interpreted effectively, yet it is also essential that their use does not fatigue or inappropriately distract the practitioners using them. In particular, as was highlighted by a number of participants in the interviews, the sound should not be distracting from other tasks, unless important monitoring information is being conveyed. The potential advantages of sonification systems with such properties extended to use on night shifts, in which practitioners may be less attentive and may therefore benefit from an extra cue to prompt them to notice security events.

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

Some participants stated that they had not experienced fatigue, or that they might experience fatigue if using the sonification for longer, but had not for the length of this study. Others stated that using the sound in the study had been fatiguing, and suggested design alterations to reduce fatigue centred around reducing the overall level of noise. These changes involved either altering the data-parsing method to include thresholds or filtering, or altering the sound engineering (improving the sound of the instruments, for example). The limitations to our study design – the time pressure, and the real-time presentation of information in the SIEM tools (which were described as unrealistic by a number of participants) – may have contributed to their experience of fatigue. As such, our results cannot constitute a reliable assessment of whether the use of sonification would cause any additional fatigue to practitioners carrying out their work in SOCs.

Issues may arise relating to the scalability of the sonification approach. Participants in the interview described the large quantities of information that need to be represented for network monitoring: both in terms of the huge volumes of network data, and of the many aspects of information (such as the various protocols and applications used in communications) that need to be represented. The large number of monitoring devices used in SOCs about which information might need to be represented was cited as another potential scalability issue. As a potential way of improving scalability, it was suggested that it may be beneficial if participants could configure sonification systems to represent only those aspects of the data they are interested in knowing about (so that only packets sent over user-selected protocols or applications are represented, for example).

VIII. Conclusion and Future Work

The aim of this article was to explore the extent to which using sonification could aid in SOC working practice. We addressed this aim through a study in which security practitioners with experience of working in SOCs carried out a scoped set of tasks relevant to SOCs using both a SIEM tool, and a SIEM tool that incorporated sonified information.

Our results showed that participants using the sonification alongside the SIEM when monitoring as a non-primary task (whilst carrying out a separate primary task) detected network traffic and alerts following an FTP brute-force attack more efficiently and identified its nature more accurately, as well as detecting the alerts produced by an IDS more accurately and efficiently. Furthermore, participants monitoring the network security as a single primary task detected alerts more accurately and efficiently when using the sonification. The precision of the anomalous-traffic detections made by participants multitasking using the sonification was lower, however: more false-positive detections were made by participants (according to our analysis of network-traffic awareness), suggesting that participants may have been misled by the sound.

We also gathered and analysed the spoken views of participants on various topics, including their experience using the study tools, the potential utility of sonification to their work, and the study design. This highlighted requirements around the design of sonification systems to enable their integration into SOCs, and showed a division in the views of participants on the contexts in which sonification systems may aid in their own work in the SOC, and in SOC working practice in general. We discussed the various limitations of our study, and of short-term studies in which resources are designed by the researchers in general, for providing realistic assessments of the utility of new network-security monitoring approaches for SOCs, and qualified our findings in light of these limitations.

As future work, it is important that longitudinal studies are run in which sonification systems are deployed in operational SOCs and their effects studied, in order to understand the true utility of the approach and the severity of the challenges to integrating sonification systems into this environment. Future research should also explore the balance between presenting information to security practitioners, and automating processes: while the role of practitioners is key, it may be advantageous to perform certain tasks automatically. Studies that compare the advantages of the various data-presentation approaches (e.g., visualization, sonification) and automated approaches (e.g., signatures and machine learning) for use in SOC tasks could help ascertain how these approaches can be used in an effective and complementary way.

Acknowledgement

The authors would like to thank the Engineering and Physical Sciences Research Council (EPSRC) for funding this work.

References

[1] CIC-IDS Intrusion Detection Evaluation Dataset. http://www.unb.ca/cic/datasets/ids-2017.html [Accessed 25-May-2018].

[2] Louise Axon, Bushra Alahmadi, Jason R C Nurse, Michael Goldsmith, and Sadie Creese. Sonification in security operations centres: what do security practitioners think? In *Proceedings of the Workshop on Usable Security (USEC) at the Network and Distributed Systems Security Symposium (NDSS)*, 2018.

[3] Louise Axon, Michael Goldsmith, and Sadie Creese. Sonification mappings: estimating effectiveness, polarities and scaling in an online experiment. *Journal of the Audio Engineering Society*, 2018. (to appear).

[4] Louise Axon, Jassim Happa, Michael Goldsmith, and Sadie Creese. Hearing attacks in network data: an effectiveness study. *Computers & Security*, 83:367–388, 2019.

[5] Louise Axon, Jason R C Nurse, Michael Goldsmith, and Sadie Creese. A formalised approach to designing sonification systems for network-security monitoring. *International Journal on Advances in Security*, 10(1&2):26–47, 2017.

[6] Balaji Balakrishnan. Security Data Visualization, 2013. https://www.sans.org/reading-room/whitepapers/metrics/security-data-visualization-36387 [Accessed: 05-July-2018].

[7] M. Ballora, N. Giacobe, and D. Hall. Songs of cyberspace: an update on sonifications of network traffic to support situational awareness. In *SPIE Defense, Security, and Sensing*, pages 80640P–80640P. International Society for Optics and Photonics, 2011.

[8] Complex Data Visualized. Security Log Visualization with a Correlation Engine, 2013. complexdatavisualized.com/tag/siem [Accessed: 05-July-2018].

[9] Anita D'Amico, Kirsten Whitley, Daniel Tesone, Brianne O'Brien, and Emilie Roth. Achieving cyber defense situational awareness: A cognitive task analysis of information assurance analysts. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 49, pages 229–233. SAGE Publications, 2005.

[10] Mohamed Debashi and Paul Vickers. Sonification of network traffic flow for monitoring and situational awareness. *PloS one*, 13(4):e0195948, 2018.

[11] Mohamed Debashi and Paul Vickers. Sonification of network traffic for detecting and learning about botnet behaviour. *IEEE Access*, 2018.

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2019.2931557, IEEE Transactions on Dependable and Secure Computing

[12] Andy Field. *Discovering statistics using IBM SPSS statistics*. Sage, 2013.

[13] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European Conference on Information Retrieval*, pages 345–359. Springer, 2005.

[14] T. Hermann, A. Hunt, and J. Neuhoff. *The sonification handbook*. Logos Verlag Berlin, 2011.

[15] Tobias Hildebrandt, Thomas Hermann, and Stefanie Rinderle-Ma. Continuous sonification enhances adequacy of interactions in peripheral process monitoring. *International Journal of Human-Computer Studies*, 2016.

[16] Nigel King. Template analysis. In G. Symon and C. Cassell, editors, *Qualitative Methods and Analysis in Organisational Research: A Practical Guide*, pages 118–134. Sage Publications Ltd, Thousand Oaks, CA, 1998.

[17] Ben Schneiderman. Eight golden rules of interface design. *Disponible en*, 1986.

[18] Iman Sharafaldin, A Habibi Lashkari, and Ali A Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *Proceedings of fourth international conference on information systems security and privacy, ICISSP*, 2018.

[19] Sathya Chandran Sundaramurthy, Alexandru G Bardas, Jacob Case, Xinming Ou, Michael Wesch, John McHugh, and S Raj Rajagopalan. A human capital model for mitigating security analyst burnout. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 347–359, 2015.

[20] Sathya Chandran Sundaramurthy, Jacob Case, Tony Truong, Loai Zomlot, and Marcel Hoffmann. A tale of three security operation centers. In *Proceedings of the 2014 ACM Workshop on Security Information Workers*, pages 43–50. ACM, 2014.

[21] Sathya Chandran Sundaramurthy, Michael Wesch, Xinming Ou, John McHugh, S Raj Rajagopalan, and Alexandru G Bardas. Humans are dynamic-our tools should be too. *IEEE Internet Computing*, 21(3):40–46, 2017.

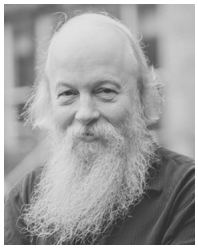
[22] Bruce N Walker. Consistency of magnitude estimations with conceptual data dimensions used for sonification. *Applied Cognitive Psychology*, 21(5):579–599, 2007.



Louise Axon is a Research Associate in Cybersecurity at the University of Oxford. She received a BA degree in Mathematics and Music from Cardiff University in 2013, an MSc in Mathematics of Cryptography and Communications from Royal Holloway, University of London in 2014, and a DPhil in Cybersecurity from the University of Oxford in 2019. Her DPhil research explored the design and utility of sonification systems for network-security monitoring tasks, and was funded by the EPSRC.



Sadie Creese is Professor of Cybersecurity in the Department of Computer Science at the University of Oxford. She was founding director of the Global Cyber Security Capacity Centre at the Oxford Martin School and a member of the Coordinating Committee for the Cyber Security Oxford network. She is engaged in a broad portfolio of cybersecurity research spanning situational awareness, visual analytics, risk propagation and communication, threat modelling and detection, network defence, dependability and resilience and privacy.



Michael Goldsmith is Director of the Global Cyber Security Capacity Centre at the Oxford Martin School and a Senior Research Fellow in the Department of Computer Science and at Worcester College, University of Oxford. His research spans a wide range of topics within security, from the mathematical to the social. He received his DPhil in Computation from Oxford University three decades ago for work on support for specification logics, and has also worked in concurrency theory and formal verification through exhaustive state-exploration.



Jassim Happa is a Research Fellow in the Department of Computer Science at the University of Oxford. His research topics include: Computer Graphics, Cybersecurity, Human Factors, Resilience and Visualization. He obtained his BSc (Hons) in Computing Science at the University of East Anglia in 2006. After a year of working as an Intrusion Detection System (IDS) analyst, he began his PhD in Engineering at the University of Warwick in 2007 in the topic of computer graphics, which he defended in 2012. He has worked at Oxford since 2011 .



Alastair Janse van Rensburg is a Research Associate in Cybersecurity at the University of Oxford. He received an MMath in Mathematics from the University of Warwick in 2014, and is currently completing a DPhil in Cybersecurity from the University of Oxford. His DPhil research involves the application of graph theory to network security, in particular looking at how attack graphs can be applied in practical situations.

1545-5971 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.