EEG-based Evaluation of Cognitive Workload Induced by Acoustic Parameters for Data Sonification

Maneesh Bilalpur
CVIT, IIIT Hyderabad
India
maneesh.bilalpur@research.iiit.ac.in

Mohan Kankanhalli
School of Computing, National University of Singapore
Singapore
mohan@comp.nus.edu.sg

Stefan Winkler
Advanced Digital Sciences Center, University of Illinois at Urbana-Champaign
Singapore, Singapore
stefan.winkler@adsc-create.edu.sg

Ramanathan Subramanian
Advanced Digital Sciences Center, University of Illinois at Urbana-Champaign
Singapore, Singapore
ramanathan.subramanian@ieee.org

ABSTRACT

Data Visualization has been receiving growing attention recently, with ubiquitous smart devices designed to render information in a variety of ways. However, while evaluations of visual tools for their interpretability and intuitiveness have been commonplace, not much research has been devoted to other forms of data rendering, *e.g.*, sonification. This work is the first to automatically estimate the cognitive load induced by different acoustic parameters considered for sonification in prior studies [9, 10]. We examine cognitive load via (a) perceptual data-sound mapping accuracies of users for the different acoustic parameters, (b) cognitive workload impressions *explicitly* reported by users, and (c) their *implicit* EEG responses compiled during the mapping task. Our main findings are that (i) low cognitive load-inducing (*i.e.*, more intuitive) acoustic parameters correspond to higher mapping accuracies, (ii) EEG spectral power analysis reveals higher α band power for low cognitive load parameters, implying a congruent relationship between explicit and implicit user responses, and (iii) Cognitive load classification with EEG features achieves a peak F1-score of 0.64, confirming that reliable workload estimation is achievable with user EEG data compiled using wearable sensors.

KEYWORDS

Data Sonification, EEG, Cognitive Workload, Acoustic parameters.

1 INTRODUCTION

Understanding and sensemaking from multi-dimensional data is a challenge, since the traditional medium for visual data representation and communication is typically restricted to two or three dimensions. This calls for visualization and content delivery tools utilizing alternative sensing mechanisms such as auditory [14], tactile [16], gustatory and olfactory [22]. Also, given the significant proportion of visually challenged persons the world over (around 253 million people are visually impaired[1]), employing purely visual communication techniques makes information inaccessible to a large section of the society.

Nevertheless, attempts to use non-visual modalities for encoding data attributes are few. Among these, data sonification, where data attributes are conveyed via psychoacoustic signals, is a relatively mature technique with applications in multiple fields such as astrophysics and neurology [10]. Two recent works that investigate the suitability/ease-of-understanding of data-to-sound mappings are [9] and [10]. In [9], a study evaluating auditory parameters (such as pitch, roughness, noise and sharpness) that best convey the focus level of an astronomical image is presented. This study is further extended in [10], where the perceptual congruence between three data attributes, namely, *stress*, *error* and *danger* and the aforementioned acoustic parameters is evaluated. Both evaluations are based on explicit user assessments acquired via the mouse and keyboard.

A limiting factor in real-life situations where visualizations are put to use is that explicit user feedback may not be available for improving or adapting the data rendering methodology. Acquiring user feedback via *implicit* means would therefore be critical for optimal information communication. Neuroergonomics, which examines human factors by employing neuroscientific methods, presents a viable alternative in this regard. This approach is also attractive with the advent of light-weight, wearable sensors. A number of works have explored *cognitive sensing* of users presented with visual information– the user's level of (dis)comfort with the presented information is gauged via eye movements [12, 23] or neural activity in the form of EEG [1, 27] and fNIRS [19]. Such studies have however not been performed, to our knowledge, for auditory perception tasks.

This paper builds on prior works [9, 10], and seeks to estimate the cognitive load of users mapping acoustic parameters to data attributes via implicit means. Cognitive load induced by different auditory parameters is examined via explicit as well as implicit user responses– participants completed (a

Accepted for publication in the proceedings of the 20th ACM International Conference on Multimodal Interaction, Colorado, USA.
DOI:10.1145/3242969.3243016

[1] http://www.who.int/en/news-room/fact-sheets/detail/blindness-and-visual-impairment

subset of questions from) the NASA-TLX questionnaire [11] to convey their impressions regarding how *easy* (low cognitive load condition) or *difficult* (high cognitive load condition) it is to map the focus level of an astronomical image to an acoustic parameter; their EEG responses while performing the mapping task were also recorded via the commercial *Emotiv* device. The main findings of this work are the following: (1) Perceptual (focus level to acoustic data) mapping accuracies are higher for parameters inducing low cognitive load as per explicit user impressions; (2) On performing power spectral analysis for the different EEG spectral bands, higher α band power is noted for low cognitive load inducing parameters, implying a congruence between explicit and implicit user responses; (3) On segregating acoustic parameters as *low* or *high* cognitive load inducing based on user impressions, we attempted cognitive load classification for each auditory presentation trial by examining the EEG data. Experiments confirm that a maximum F1-score of 0.64 is achievable with a CNN based classifier. This implies that user perception of acoustic data can be gauged via implicit means. Overall, this work makes the following contributions.

i. To our knowledge, this is the first work to expressly investigate the cognitive load induced by multiple acoustic parameters via explicit and implicit means. Our analysis combines examination of the (a) perceptual mapping accuracies for data sonification, (b) user cognitive load impressions, and (c) user EEG responses.

ii. We demonstrate congruence between explicit user impressions and implicit neural activity, Acoustic parameters inducing low cognitive load are found to be associated with higher EEG α power, in line with previous findings [27].

iii. We show that better-than-chance cognitive load classification is possible by examining the user EEG signals. A maximum F1-score of 0.64 is obtained for low-vs-high cognitive load classification using a deep convolutional neural network (CNN) classifier.

The rest of the paper is organized as follows. Section 2 motivates our work in the context of available literature, and highlights its novelty. Section 3 details the experimental design and protocol. Section 4 examines explicit user data in terms of perceptual mapping accuracies and cognitive load impressions, while Section 5 presents an analysis of the EEG data. In Section 6 we introduce the cognitive load classification experiments and discuss the results, while Section 7 concludes the paper.

2 RELATED WORK

This section reviews literature on data sonification and cognitive load estimation to motivate the need for our study.

2.1 Visualization and Data Sonification

Information Visualization (InfoViz) concerns the design and development of interactive and graphical representations of information. Interactive visualizations typically convey visual and data patterns utilizing rendering techniques that best cater to human perception and cognition [5, 20]. Among the various visualization techniques, sonification involves the use of non-speech based audio signals to convey information. A sonification system conveys data values by manipulating acoustic parameters such as sound frequency (pitch) or tempo. Two recent works that attempt to identify the optimal acoustic parameters for conveying different (types of) data attributes are [9] and [10].

A study presented in [9] explores the utility of *sharpness*, *roughness*, *noise*, *pitch* and a combination of *roughness* and *noise* for conveying the focus level of an astronomical image. The sound parameters are carefully chosen, upon reviewing sonification literature in great detail. Sound parameters in [9] are evaluated by: (1) computing the mean perceptual data:sound mapping accuracy for focus level determination, and (2) comparing the performance of the sound parameters against the visual stimuli to evaluate if the acoustic parameters can indeed serve as effective proxies for conveying the visual information. The study concludes with two main findings: (a) Acoustic parameters that converge on a clear/pure sound are optimal for focus determination, and (b) The investigated auditory parameters can provide effective substitution for visual information.

An extension of the above study is presented in [10]. Here, the ability of parameters such as *roughness*, *noise* and *pitch* for conveying negative data attributes such as *error*, *danger* and *stress* is explored. This study concludes that the effectiveness of sound parameters is governed by the ease with which users can perceive the data:sound mapping. An earlier study detailed in [26] suggests that intuitive data:sound mappings may not result in the best user performance in terms of accuracy or response times.

2.2 Cognitive Workload Estimation

Cognitive workload (or mental workload) has been traditionally employed as a standard for measuring task difficulty by many previous works [1, 2, 27]. In particular, there is a large body of work correlating how neural activity captured via EEG [2, 27] and fNIRS signals [19] can enable cognitive workload assessment. The advantage of employing neural signals such as EEG and fNIRS is that they can be captured via light-weight and wearable commercial devices, as against other physiological modalities that require data to be recorded with bulky and specialized lab equipment.

The utility of EEG for measuring cognitive workload has been demonstrated by many previous works [2, 27]. The observation that changes in the EEG θ and α band power are indicative of memory load is made in [27]. Specifically, low cognitive load is found to be associated with higher α band power. The cognitive load induced by visualizations such as bar and quartile plots is studied via EEG analysis in [1]. NASA-TLX parameters are employed to explicitly obtain cognitive load ratings from users, and the authors demonstrate prefrontal cortex activity captured during task performance is relevant to the working memory consumption.

Deep learning for cognitive load estimation is proposed in [2]. The ability of convolutional neural networks (CNNs) to preserve spatial, spectral and temporal structure of EEG is exploited in this work. Spectral band maps are first synthesized from EEG data corresponding to low and high mental workload tasks, and automated classification of low/high cognitive load from EEG data is then attempted.

2.3 Analysis of Related Work

Upon reviewing related literature, one can note that (1) The evaluation of auditory parameters for data rendering is still an active area of research. While the authors of [9] and [10] base their findings on perceptual mapping accuracies obtained from user responses, a direct estimation of user cognitive load is not attempted in either of these works; (2) While mental workload estimation has been explored via cognitive sensing as users perform visual processing, to the best of our knowledge we have not encountered equivalent studies for sonification.

In this regard, we present the first work towards estimating cognitive load of users by analyzing their EEG signals. A salient aspect of our work is that these EEG signals are captured via a light-weight and wearable *Emotiv* device. It is reasonable to expect that users will be willing to use such sensors to feedback their cognitive state as they perform real-life perception tasks. The next section details the stimuli and protocol employed for estimation of sonification-induced cognitive workload.

3 STIMULI AND EXPERIMENTAL DESIGN

For the purpose of this study, we used the acoustic parameters employed in [9], namely, *noise*, *pitch* (pure sinusoidal tones in a C-major scale as in Experiment 2 of [9]), *roughness* and *combination* of roughness and noise. Furthermore, we used the original astronomical images with varying focus levels for comparison. In this section, we describe the stimuli used and the adopted experimental protocol.

3.1 Stimuli

3.1.1 Visual. For benchmarking purposes, we used the astronomical images representing varying levels of focus in our study. These images are presented in Figure 1. A color image showing a telescopic view of the M110 galaxy was used for the highest focus level (level 10). Blurred versions of this image were generated using OpenCV via smoothing with different kernel sizes. The visual images are found to convey the focus level most precisely to users in [9].

3.1.2 Noise. Similar to [9], we used a 1000 Hz pure tone and broadband white noise to convey 10-levels of varying focus. As suggested in [9], we exploit the possible association of noise with a negative attribute such as blur to denote the image focus level. A perfectly focused image (focus level of 10) corresponds to a 100% pure tone while a 1-focus level (*i.e.*, completely de-focused) image corresponds to 100% noise.

3.1.3 Pitch. The sensation of acoustic frequency is referred to as Pitch. This parameter is usually found in music. Pure sinusoidal tones in a C-major scale (plus 2 extra notes to make the range 10 notes long) beginning at the middle C (C4, freq = 261.63 Hz) and ending on E6 (freq = 1318.51 Hz) were used. Loudness variations for the various frequencies were compensated for. Higher frequencies denoted higher focus levels.

3.1.4 Roughness. Having demonstrated in [9] the drawback of using a noisy carrier signal for roughness representation, we used a 100% pure-tone that was amplitude modulated with 0, 2, 4, 7, 11, 16, 23, 34, 49 and 70 Hz. We believe that the dissonance in the acoustics best represents image blur. Hence we used a pure tone to represent the image of focus level-10.

3.1.5 Combined Roughness and Noise. This is the second best performing acoustic parameter in [9] after the visual images. Thus, we used this direct pairing of roughness and its corresponding noise. Analogous to the above described individual parameters, a pure-tone was used to represent the highest focus level of the image and the combination of 100% noise modulated by 70Hz corresponds to focus level-1.

3.1.6 Combined Image plus Acoustics. In addition to the above mentioned parameters, we also used a mix of *visual* images *combined* with roughness and noise to explore if a combination of the visual and acoustic modalities enabled better identification of the image focus level. This parameter is referred to as *Visual*-plus-*Combined* in the remainder of the paper.

3.2 Cognitive Workload parameters

Cognitive Workload is a complex construct of intrinsic, extrinsic and germane load [8]. Hence, we utilized the multidimensional rating design of NASA-TLX [11] to acquire explicit cognitive load impressions from users. We employed a subset of the parameters to rate the task difficulty, in particular the three factors of *effort* (Rate the level of effort you needed to put in to complete the task on a scale of 0 (lowest)–4 (highest)), *mental demand* (Rate the level of stress you endured while performing the task on a 0–4 scale) and *frustration* (Rate the level of frustration you experienced while performing the task on a 0–4 scale). We hypothesized that these three parameters had complementary contributions to the user workload, given the nature of the task.

3.3 Participants

20 participants (16 male) with an average age of 28.9 ± 4.9 years took part in our study. None of them had any formal training in music. Users had to perform the experiment for about one hour, and were financially compensated for their participation. The experimental design was approved by the local ethics committee.

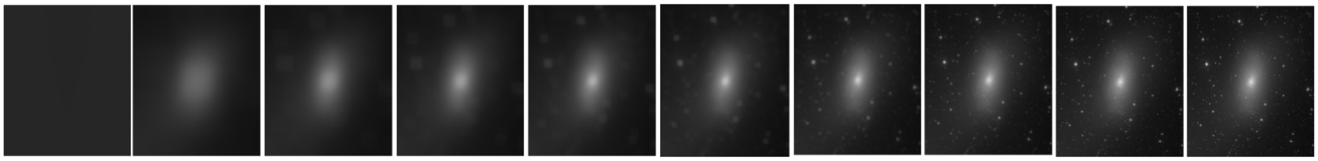

Figure 1: Astronomical images shown with various focus levels. The image focus level progressively increases from 1–10 from left to right.

3.4 Protocol

The experiment was conducted in two sessions, where each session was divided into six blocks. Each of the six blocks corresponded to one of the visualization parameters (4 acoustic, 1 visual and 1 visual-plus-acoustic). Within each block, the 10 stimuli were played for 2 sec each in random order and repeated thrice, leading to a total of 30 trials (stimulus presentations) in a block.

Session 1 involved playing of a stimulus following which, the user had to immediately rate the corresponding focus level on a scale of 1–10. Session 1 therefore involved **Immediate Recall** (IR) from the user. In Session 2, the 10 stimuli corresponding to a particular visualization parameter were again repeated thrice in random order, but at the beginning of each of the three repetitions, users were instructed to click a 'Yes' radio button if they inferred that the played stimulus corresponds to a particular focus level. The focus level of interest was pre-specified prior to each of the three repetitions, and the objective here was to facilitate perceptual comparisons between the stimuli presented over successive trials to arrive at a decision. Session 2 is similar to the rapid serial visual presentation (RSVP) protocol adopted in psychophysical studies, and is termed the **Compared Recall** (CR) session. To enable at least one comparison by the user, we ensured that the target focus level stimulus was not rendered in the first trial following the target focus level specification.

In the IR session, each stimulus was presented for 2s, preceded by a a fixation cross which was displayed for 500ms. The stimulus was followed by a question to let users ascertain the corresponding focus level on a 1–10 scale. A 10s timer was displayed on the screen during the judgement task, and the next stimulus was automatically presented if the user failed to respond within the 10s time-frame. All parameters remained identical in the IR and the CR sessions with the exception of each stimulus being followed by a 3s response time in CR (as against a 10s time-frame in IR).

Upon completion of each block, users were required to rate their experience on the subset of NASA-TLX parameters (mentioned above) to evaluate the cognitive workload. No technical terms were used to represent the various stimuli, and participants were instructed regarding the stimulus type through a coded representation (Type 1–6). Users were also made to perform a practice session before each session, where they were familiarized with all the stimuli employed in the experiments.

3.5 Data Acquisition

As users performed the focus level detection task, we acquired their neural responses (which we hypothesized to capture the cognitive workload experienced) via the 14-channel consumer-grade *Emotiv Epoc* EEG device. Both the IR and CR sessions were split into two halves to calibrate the device regularly, and to prevent participants from experiencing fatigue. The *Epoc* device used has a 128 Hz sampling rate. Our experimental protocol was designed using Matlab PsychToolbox [6].

4 EXPLICIT RESPONSE ANALYSIS

To begin with, we attempt to model cognitive workload in terms of data:parameter mapping (or recognition) accuracies, and cognitive workload impressions provided by users in terms of the NASA-TLX ratings.

We analyzed the accuracy of user responses for the various stimuli presented and their combinations. We summarize and compare the results with prior work described in [9]. As with [9], we considered a 10% error margin while computing accuracies.

4.1 Recognition Rate

For Immediate Recall, the highest mapping accuracy was observed for *visual*-plus-*combined* (0.83 ± 0.11) while the least accuracy was observed for *roughness* (0.48 ± 0.19) (see Figure 2(left)). This implies that the combination of visual plus acoustic cues facilitated better inference of the image focus level by users. The acoustic stimuli followed a similar trend as in [9] with the *combined* rendering of *roughness* and *noise* (accuracy of 0.71 ± 0.18) being the easiest to map. The only exception in our study is that *pitch* outperforms *roughness.* The *visual* stimuli (accuracy = 0.78 ± 0.18) were easier to perceive than all acoustic parameters, in line with one's expectations.

For Compared Recall, very similar mapping accuracies were noted for *visual*-plus-*combined* (0.82 ± 0.33) and *visual* (0.84 ± 0.22) (see Figure 2(left)). Among acoustic parameters, *Noise* (0.67 ± 0.26) and *Combined* (0.64 ± 0.32) were most suitable for detecting the target focus level, while *pitch* was the most difficult (0.54 ± 0.3). Overall, a similar trend as in [9] was observed. Higher standard deviations in the CR session suggest that focus level identification based on stimulus comparison was more difficult than individually mapping each presented stimulus as in the IR protocol.

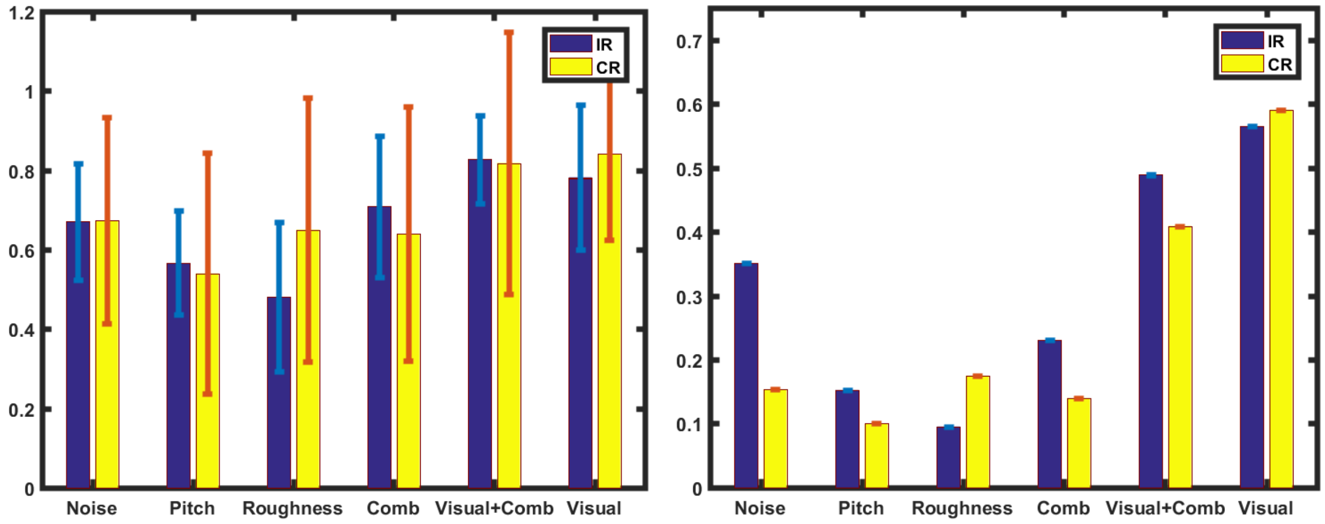

Figure 2: (left) Mapping accuracy with the different parameters. Error-bars denote unit standard deviation. (right) Proportion of *low* cognitive load trials for each parameter.

4.2 Cognitive Workload Ratings

We considered the mean of the three NASA-TLX parameters (*effort*, *mental demand* and *frustration*) to evaluate the overall mental workload. The average score was thresholded at the mean value of 2 (since the used scale was 0–4) to quantize or characterize a parameter block as inducing *low*/*high* workload. To examine the correlation between recognition rates and the cognitive workload impressions for a given visualization parameter, we computed the proportion of low cognitive workload trials for each of the six parameters. We hypothesized that a visualization parameter that induces low cognitive workload should correspond to higher mapping accuracy, *i.e.*, parameters that correspond to higher accuracies in Figure 2(left) will also have many low cognitive load blocks as shown on Figure 2(right). To this end, we computed Pearson correlation coeffcients between the mapping accuracies and the proportion of low cognitive load blocks. For IR, we observed a correlation $\rho = 0.8339$, $p<0.05$, while a $\rho = 0.8631$, $p<0.05$ was noted for CR. The above correlations confirm that there exists a congruent relationship between the recognition rates and cognitive workload.

Given that the ultimate objective of this study is to assess mental workload via the implicitly acquired EEG data, we need to label *trials* (denoting the presentation of an individual stimulus) as *high* or *low* cognitive load inducing. To this end, labels were generated for the IR and CR sessions based on user workload impressions. As seen from Figure 2(right), we categorized *pitch*, *roughness* and *combined* as *high* load parameters for IR, while *noise*, *visual*, and *visual* plus *combined* as *low* load parameters. This categorization was based on the maximum inter-parameter difference of 0.12 noted between the *combined* and *noise* parameters in Figure 2(right). Similarly for CR, the high load parameters included *pitch*, *combined*, *noise* and *roughness* while the low load parameters were *visual*, and *visual* plus *combined*. We observed that the acoustic parameters were rated to induce relatively higher load than the *visual* or *visual* plus *combined* parameters as anticipated.

5 EEG ANALYSIS

In this section, we describe the data preprocessing techniques and our inferences from the EEG data acquired during the memory workload evaluation task.

5.1 Data Preprocessing

The EEG data acquired using the consumer-grade *Epoc* device is highly susceptible to external electrical noise, motor activity (apart from the task-related activity) such as eye-blinks, head and muscle movements. Thus, the EEG data is subjected to a preprocessing pipeline to eliminate artifacts.

We extracted epochs of 2.5s duration for each trial (comprising 0.5s of fixation presentation and 2s of stimulus presentation). We performed removal of the baseline neural activity DC-offset using the EEG response for the 0.5s fixation duration. Further, the EEG signals were subjected to (a) band-limiting between 0.1–45 Hz, (b) visual rejection of noisy epochs and (c) Independent Component Analysis-based rejection of artifacts corresponding to eye-blinks and movements. Muscle movement artifacts are removed upon band limiting EEG as they are chiefly concentrated in the 40–100 Hz band, and via manual removal of noisy ICA components.

5.2 ERP analysis

Several previous works have investigated Event Related Potentials (ERPs) for human-centric decision tasks like emotion recognition [4], image annotation [17] and error identification [25]. These markers act as a bridge between neuroscience and behavioral studies, facilitating better understanding of the

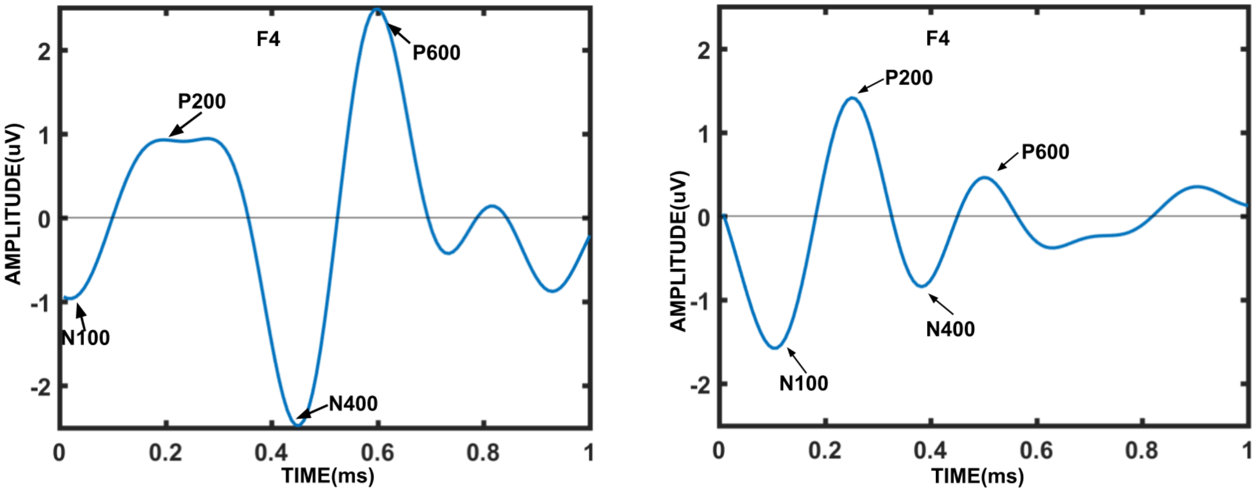

Figure 3: ERP plots for F4 channel corresponding to *immediate recall* (left) and *compared recall* (right) sessions, with different ERP components labeled.

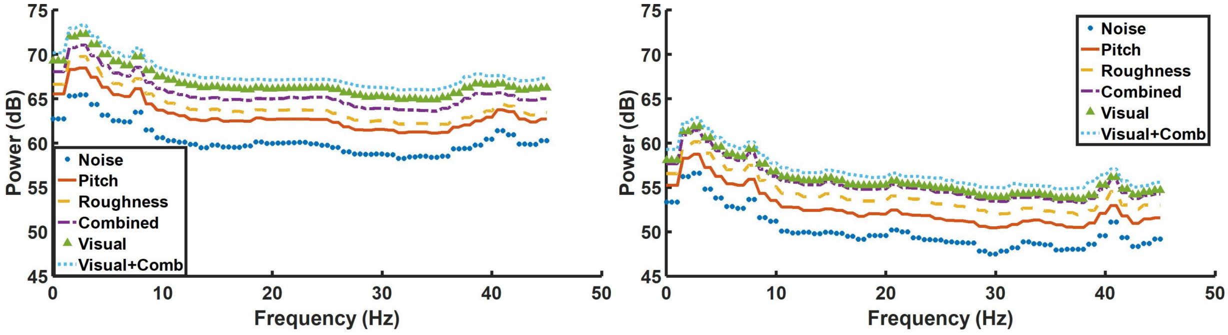

Figure 4: Power spectra plots for the different visualization parameters in the IR (left) and CR (right) tasks.

structural and functional relevance of neural activity to the task on hand, and validation of the EEG data.

Figure 3 shows the ERP components observed upon analyzing EEG data from all IR and CR trials. In particular, we emphasize the existence of N100 and P200 components in both the IR and CR sessions. The N100 component is known to reflect processing of acoustic cues, and is typically followed by the P200 component which is fronto-centrally distributed [18]. Also, amplitude of the N100 component is known to be sensitive to the level of attention [3]. We attribute higher N100 amplitude for CR to the experimental design, as the CR protocol forces users to compare the current stimulus against prior ones for target detection, thereby demanding greater attention than the IR protocol (as also reflected via higher variance in mapping accuracies for CR). In addition to the N100 and P200 components, we also encountered the N400 and P600 components in the F4 channel. Several works [7, 15] confirm the presence of N400, P600 components in language comprehension and semantic memory understanding tasks. However, their presence while processing visualizations has not been explored as yet.

5.2.1 Spectral Analysis. Prior studies on memory workload for the n-back task [27] have demonstrated the existence of differences in α (8-13 Hz), θ (4-7 Hz), lower β (13-16 Hz) and higher γ (40-45 Hz) EEG band powers with varying task difficulty. We investigated the power spectrum averaged over all the EEG channels for any existence of such cues. The power spectrum analysis (Figure 4) suggests an increased activity in the α band (as well as other frequency bands such as δ (1-4 Hz) and γ) for low workload parameters identified from Section 4.2. These patterns are better observable for the CR task as compared to IR. Differences noted in [27] for the θ band are not observed in our study. It is also interesting to note that the power spectral trends for both the IR and CR tasks are similar indicating that memory workload is reliably captured by our study. Some similarities can be noted between Figure 2(right) and Figure 4. Those parameters associated with *low* cognitive workload (such as *Visual* and *Visual* plus *Combined*) are associated with higher spectral power (in particular the α band) than *high* workload parameters (such as *pitch* and *noise*).

6 COGNITIVE LOAD CLASSIFICATION

This section describes how we train classifiers from the compiled EEG data and cognitive workload labels to implicitly assess memory workload from EEG. To this end, we

Table 1: Convolutional neural network parameters.

Parameter	Value
Learning rate	0.01
Kernel size	3
Stride size	2
Pool size	2
Batch size	32
# Kernels(layer wise)	16,32,32
Momentum	0.9
Weight decay	0.0001
Dropout	0.1

labeled all trials corresponding to parameters associated with *low*/*high* workload (from Section 4.2) for the IR and CR tasks accordingly, *i.e.*, *Visual*-plus-*Combined*, *Visual* and *Noise* parameters were considered to be low workload inducing for IR, while only *Visual* and *Visual*-plus-*Combined* were deemed as low workload inducing for CR.

6.1 Classification methods

6.1.1 Traditional methods. We employed traditional machine learning methods– Naive Bayes (NB), Linear Discriminant Analysis (LDA), SVM with linear kernel (LSVM), and SVM with RBF kernel (RSVM) for EEG-based *low*/*high* workload categorization. Prior to classification, the 14 channel, 2s long stimulus-duration data was vectorized and subjected to dimensionality rejection via principal component analysis to retain 90% variance. We performed 10 repetitions of 5 fold cross validation on the data. Given the class imbalance for the IR and CR tasks, we chose F1-score as the metric to evaluate classifier performance.

6.1.2 Convolutional Neural Network. The CNN feature learning can achieve better dimensionality reduction by projecting the EEG data onto a robust feature space for preserving task relevance and invariance to noise.

We adopted a 3-layer CNN model proposed in [21] to learn a robust representation from EEG data for classification of cognitive workload. Three convolution together with rectified linear unit (ReLU) activation function and average pooling layers are stacked to extract task specific features (Figure 5). The convolutions employed are 1-dimensional along time. Batch normalization [13] is used after the third CNN layer to reduce the internal covariate shift and accelerate the training. To prevent over-fitting, we used dropout after the fully connected layer with 128 neurons. We finally classify with *softmax* over 2 output neurons. The number of kernels increase with the depth of the convolution network as analogous to the VGG architecture [24]. We have optimized the network for categorical cross entropy using vanilla stochastic gradient descent with Nestrov momentum and weight decay. CNN hyper-parameters are specified in Table 1. The values for these hyper-parameters are mainly adopted from [21] or otherwise decided by cross-validation (10 repetitions of 5-fold).

6.2 Results

Table 2 summarizes the performance of various classifiers in terms of F1-scores ($\mu \pm \sigma$) obtained over 50 runs. Precision and recall values are also tabulated for better insights. The obtained results clearly show that the LDA, LSVM and CNN classifiers outperform Naive Bayes and RSVM, and achieve better-than-chance cognitive load classification. The fact that LSVM outperforms RSVM for both IR and CR tasks implies that non-linear kernels do not result in better classification performance for our data. Traditional classifiers result in low precision for the IR task, and low recall for the CR task. Overall, the CNN classifier achieves the most balanced performance in terms of precision and recall, and produces the best workload classification performance for the IR task (F1 = 0.64). Nevertheless, its performance decreases for the IR task presumably because of the class imbalance and the fact that training data was far fewer for this condition (4 *low* cognitive load vs 2 *high* cognitive load parameters, and only those trials eliciting a user response were employed for training). Correspondingly, higher variance in the performance of all classifiers is noted in the CR condition. The LSVM classifier performs best for the CR task. Overall, our classification results emphasize the need for efficient and robust feature learning for workload estimation from noisy EEG signals.

7 DISCUSSION AND CONCLUSION

This work examines the efficacy of the *pitch*, *roughness*, *noise*, and *combined* acoustic (in combination with visual) parameters for conveying image focus level in terms of explicit data:parameter mapping accuracies and user cognitive load impressions conveyed via NASA-TLX attributes. Specifically, parameters associated with *low* cognitive load result in higher mapping accuracies in line with one's expectations. Furthermore, we perform automated classification of cognitive load from EEG signals acquired via a commercial wireless device, and labels based on explicit responses to achieve a maximum (and significantly above-chance) F1-score of 0.64. In terms of novelty, our work improves over evaluation studies conducted in [9, 10] which exclusively rely on explicit user responses for workload understanding.

On the whole, our findings mirror those in [9] and [10]. The best performing acoustic parameter in terms of data:sound mapping accuracies is the *combined* rendering of *roughness* and *noise* (Section 4.1), which owing to its negative attributes best conveys *image blur* (inverse of *focus level*). We observe that focus recognition accuracy is higher for visual stimuli than acoustic stimuli; nevertheless, the recognition accuracy is highest for *visual*-plus-*combined* implying that acoustic information augments visual information towards determining the focus level. Congruence between recognition accuracies and cognition load is also highly noticeable. Specifically, the *visual* and *visual*-plus-*combined* conditions are reported to induce *low* memory workload over a majority of the trials, while *noise* is observed to be the most intuitive in terms of acoustic parameters based on NASA-TLX responses.

We designed two different tasks for determining the image focus level in our experimental design– immediate recall, where users had to immediately detect focus level from the presented stimulus, and compared recall where the user needed to detect a pre-specified target level from among

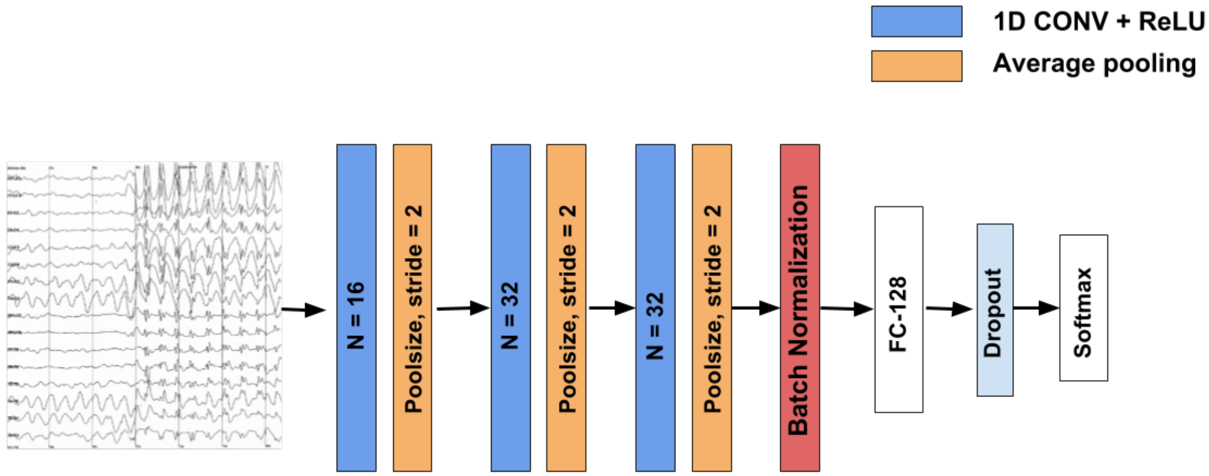

Figure 5: CNN architecture showing various layers in the model and parameters.

Table 2: Classification results summary (* denotes F1-score distribution significantly above chance level with p<0.05).

	IR			CR		
Classifier	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
NB	0.48 ± 0.01	0.48 ± 0.03	0.48 ± 0.02	0.49 ± 0.06	0.37 ± 0.06	0.42 ± 0.04
LDA	0.50 ± 0.01	0.61 ± 0.02	0.55 ± 0.02*	0.56 ± 0.05	0.50 ± 0.07	0.53 ± 0.07*
LSVM	0.48 ± 0.01	0.54 ± 0.02	0.51 ± 0.02*	0.58 ± 0.04	0.52 ± 0.07	**0.55 ± 0.06***
RSVM	0.37 ± 0.04	0.46 ± 0.15	0.41 ± 0.03	0.34 ± 0.01	0.55 ± 0.00	0.42 ± 0.00
CNN	0.64 ± 0.02	0.64 ± 0.02	**0.64 ± 0.02***	0.54 ± 0.10	0.52 ± 0.06	0.52 ± 0.07*

serially presented stimuli. We hypothesized that CR would be more facile for visualization understanding since comparisons would enable a better assessment of the rendered visualizations. However, different from our expectation, a greater variance in IR recognition accuracies suggests that focus level identification was perhaps more difficult for the IR task. It is also pertinent to note that the N100 ERP component was stronger for the IR task, and indicative of the greater cognitive attention required in this setting. A more thorough analysis of the focus levels that were easy/difficult to detect in the IR and CR settings is required as part of future work.

Power spectral analysis also confirmed that low workload parameters are characterized by higher α, δ and γ band powers as compared to high workload parameters, and that the power spectral densities for the CR task are significantly lower than those for the CR condition. In terms of binary cognitive workload classification, the CNN model produces the most balanced performance in terms of precision and recall, and results in the peak F1 score of 0.64 for the IR condition. Relatively lower F1 scores achieved for CR can be attributed to fewer training data available in this condition (about $\frac{1}{7}^{th}$ of the training data available for IR owing to the experimental design). While this work compiles data from the general population, data sonification could specifically benefit the visually impaired community, and future work would evaluate sonification techniques on such participants.

ACKNOWLEDGEMENTS

We thank Steve Brewster and Jamie Ferguson for kindly providing us with the visual and auditory data used in their experiments. This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre in Singapore Funding Initiative. Maneesh Bilalpur completed this work as a visiting student at the SeSaMe Centre in the National University of Singapore.

REFERENCES

[1] Erik W Anderson, Kristin C Potter, Laura E Matzen, Jason F Shepherd, Gilbert A Preston, and Cláudio T Silva. 2011. A User Study of Visualization Effectiveness Using EEG and Cognitive Load. In *Computer Graphics Forum*, Vol. 30. Wiley Online Library, 791–800.

[2] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. 2015. Learning Representations from EEG with Deep recurrent-Convolutional Neural Networks. *arXiv preprint arXiv:1511.06448* (2015).

[3] Alan S Bellack and Michel Hersen. 1998. Introduction to Comprehensive Clinical Psychology.

[4] Maneesh Bilalpur, Seyed Mostafa Kia, Manisha Chawla, Tat-Seng Chua, and Ramanathan Subramanian. 2017. Gender and Emotion Recognition with Implicit User Signals. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 379–387.

[5] Michael Bostock and Jeffrey Heer. 2009. Protovis: A Graphical Toolkit for Visualization. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2009).

[6] David H Brainard and Spatial Vision. 1997. The Psychophysics Toolbox. *Spatial Vision* 10 (1997), 433–436.

[7] Harm Brouwer and Matthew W Crocker. 2017. On the Proper Treatment of the N400 and P600 in Language Comprehension.

Frontiers in Psychology 8 (2017), 1327.

[8] Fang Chen, Jianlong Zhou, Yang Wang, Kun Yu, Syed Z Arshad, Ahmad Khawaji, and Dan Conway. 2016. *Robust Multimodal Cognitive Load Measurement.* Springer.

[9] Jamie Ferguson and Stephen A Brewster. 2017. Evaluation of Psychoacoustic Sound Parameters for Sonification. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction.* ACM, 120–127.

[10] Jamie Ferguson and Stephen A Brewster. 2018. Investigating Perceptual Congruence Between Data and Display Dimensions in Sonification. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 611.

[11] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology.* Vol. 52. Elsevier, 139–183.

[12] Weidong Huang. 2007. Using Eye Tracking to Investigate Graph Layout Effects. In *Visualization, 2007. APVIS'07. 2007 6th International Asia-Pacific Symposium on.* IEEE, 97–100.

[13] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167* (2015).

[14] Gary S Kendall. 1991. Visualization by Ear: Auditory Imagery for Scientific Visualization and Virtual Reality. *Computer Music Journal* 15, 4 (1991), 70–73.

[15] Marta Kutas and Kara D Federmeier. 2011. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology* 62 (2011), 621–647.

[16] Beom-Chan Lee, Hyeshin Park, Junhun Lee, and Jeha Ryu. 2007. Tactile Visualization with Mobile AR on a Handheld Device. In *International Workshop on Haptic and Audio Interaction Design.* 11–21.

[17] Viral Parekh, Ramanathan Subramanian, Dipanjan Roy, and CV Jawahar. 2018. An EEG-based Image Annotation System. In *Computer Vision, Pattern Recognition, Image Processing, and Graphics.* Springer, 303–313.

[18] Silke Paulmann. 2015. The Neurocognition of Prosody. In *Neurobiology of Language.* Elsevier, 1109–1120.

[19] Evan M M Peck, Beste F Yuksel, Alvitta Ottley, Robert JK Jacob, and Remco Chang. 2013. Using fNIRS Brain Sensing to Evaluate Information Visualization Interfaces. In *Human Factors in Computing Systems.* ACM, 473–482.

[20] Kranthi Kumar Rachavarapu, Moneish Kumar, Vineet Gandhi, and Ramanathan Subramanian. 2018. Watch to Edit: Video Retargeting using Gaze. *Computer Graphics Forum* 37, 2 (2018), 205–215.

[21] Nastaran Mohammadian Rad, Seyed Mostafa Kia, Calogero Zarbo, Twan van Laarhoven, Giuseppe Jurman, Paola Venuti, Elena Marchiori, and Cesare Furlanello. 2018. Deep Learning for Automatic Stereotypical Motor Movement Detection Using Wearable Sensors in Autism Spectrum Disorders. *Signal Processing* 144 (2018), 180–191.

[22] Nimesha Ranasinghe, Kasun Karunanayaka, Adrian David Cheok, Owen Noel Newton Fernando, Hideaki Nii, and Ponnampalam Gopalakrishnakone. 2011. Digital Taste and Smell Communication. In *Conference on Body Area Networks.* ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 78–84.

[23] Michael Raschke, Tanja Blascheck, Marianne Richter, Tanja Agapkin, and Thomas Ertl. 2014. Visual Analysis of Perceptual and Cognitive Processes. In *Information Visualization Theory and Applications.* IEEE, 284–291.

[24] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014).

[25] Chi Thanh Vi, Izdihar Jamil, David Coyle, and Sriram Subramanian. 2014. Error Related Negativity in Observing Interactive Tasks. In *Human Factors in Computing Systems.* ACM, 3787–3796.

[26] Bruce N Walker and Gregory Kramer. 2005. Mappings and Metaphors in Auditory Displays: An Experimental Assessment. *ACM Transactions on Applied Perception* 2, 4 (2005), 407–412.

[27] Shouyi Wang, Jacek Gwizdka, and W Art Chaovalitwongse. 2016. Using Wireless EEG Signals to Assess Memory Workload in the n-Back Task. *IEEE Transactions on Human-Machine Systems* 46, 3 (2016), 424–435.