Symptoms of Cognitive Load in Interactions with a Dialogue System

José Lopes
Heriot-Watt University
Edinburgh, United Kingdom
jd.lopes@hw.ac.uk

Katrin Lohan
Heriot-Watt University
Edinburgh, United Kingdom
K.Lohan@hw.ac.uk

Helen Hastie
Heriot-Watt University
Edinburgh, United Kingdom
h.hastie@hw.ac.uk

ABSTRACT

Humans adapt their behaviour to the perceived cognitive load of their dialogue partner, for example, delaying non-essential information. We propose that spoken dialogue systems should do the same, particularly in high-stakes scenarios, such as emergency response. In this paper, we provide a summary of the prosodic, turn-taking and other linguistic symptoms of cognitive load analysed in the literature. We then apply these features to a single corpus in the restaurant-finding domain and propose new symptoms that are evidenced through interaction with the dialogue system, including utterance entropy, speech recognition confidence, as well as others based on dialogue acts.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; *Natural language interfaces*;

KEYWORDS

Cognitive load, Spoken Dialogue Systems, Multi-modal Systems

ACM Reference Format:
José Lopes, Katrin Lohan, and Helen Hastie. 2018. Symptoms of Cognitive Load in Interactions with a Dialogue System. In *Workshop on Modeling Cognitive Processes from Multimodal Data (MCPMD'18), October 16, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3279810.3279851

1 INTRODUCTION

In complex time-critical situations, users of interactive systems can experience high cognitive load, which can interfere with their ability to perform a task and can reduce situation awareness. These cognitive demands could be as a result of the environment, for which they have no control (e.g. fire rescue [16]); or when other team members are not performing in collaborative situations; or as a result of the interface not being optimal, for example, information overload. An understanding of the user state, specifically their cognitive load, would enable an interactive system to adapt, giving appropriate support as per the user's cognitive burden, for example, only information that is absolutely necessary for that user at that particular time.

This study is within the context of the EPSRC funded ORCA Hub [11], whose vision it is to use teams of Robots and Autonomous Intelligent Systems (RAIS) to work on offshore energy platforms to enable cheaper, safer and more efficient working practices. The goal is that through the use of such robotic systems offshore, the need for personnel will decrease. This scenario results in interesting challenges, in terms of i) enabling operators and managers to have high situation awareness at all times; ii) allowing operators to work seamlessly in teams of humans-and-robots and give RAIS directions if necessary, for example, changing the mission goal; and iii) intervening and taking control at varying levels as needed. Finally, autonomous systems must be accountable and a lack of transparency of their actions, i.e. what they are doing and why, can reduce understanding and trust [10, 20, 35]. This builds on prior work around an interactive interface for such remote autonomous called MIRIAM (Multimodal Intelligent inteRactIon for Autonomous systeMs) [9], which provides explanations of system behaviour to increase transparency [7] and has been shown to increase situation awareness [28].

In this paper, we examine what speech and dialogue data has been collected for high cognitive load scenarios and provide a survey of the *symptoms* of cognitive load, to which a system could automatically detect and adapt. These symptoms can be classed in terms of performance-based metrics of the user/system, behavioural symptoms (e.g. eye movement, linguistic phenomena such as hesitations) and physiological symptoms (e.g EEG). Given the intrusive methods of physiological data, we concentrate of the first two of these and in particular linguistic symptoms. As well as those in the literature, we provide new and novel findings and finally provide recommendations for future data collections, for which we have planned.

2 BACKGROUND

As discussed in [4], symptoms of cognitive load can be physiological, behavioural or related to the performance in a given task. Examples of physiological symptoms of cognitive load previously reported in the literature include heart rate [15], skin conductance [32], pupil dilations [1, 13], eye blinks [22] and movement [24], and EEG [34]. Work presented by Lohan et al. [23] suggests that it might be possible to distinguish different styles of processing the same stimuli using deep machine learning approaches (e.g. LSTMs). Physiological symptoms, however, tend to be captured via intrusive methods or can pose risks to the subjects. For example, using pupil diameter as a measurement for cognitive load involves some risk from the pupilar light reflex [18]. Consequently, there has been research investigating alternative modalities, which we focus on

MCPMD'18 , October 16, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-6072-2/18/10…$15.00
https://doi.org/10.1145/3279810.3279851

here. Specifically, behavioural symptoms have been shown to be observable in the audio signal [8, 14, 17, 29, 31], syntax [6], lexical choice [16], and pragmatics including phenomena such as turn-taking [8, 14].

Previous work looked at identifying symptoms of cognitive load through data collection, specifically conditions designed to induce cognitive load. High cognitive load levels could either be induced intrinsically by increasing the difficulty of the task or by extraneous load, by adding a secondary task to the main task [26]. We discuss firstly symptoms observed by a single participant in isolation, e.g. reading. However, we are primarily interested in interaction. Finally, we discuss one study that compares both Human-Human and Human-Computer interaction (the latter inducing higher cognitive load). Results are summarised in Table 1.

2.1 Participant in Isolation (H)

A number of studies involve humans performing a task in isolation. For the interest of this paper, we will focus on studies where subjects had to speak. In the studies described in [17, 29], cognitive load is induced via a secondary task, while reading aloud a text extract and answering questions about its content [16] or performing a logical reasoning test [29]. In [17], the durations for both silent and filled pauses were higher in the high cognitive load condition, as well as the values for latency. In [29], speech rate was significantly higher, F0 was marginally higher F0 and the energy decay was significantly lower for subjects in the high load condition. In [14], the main task was to formulate questions according to pictures shown in a graphical interface featuring a virtual airport. Three methods were used to induce cognitive load: a secondary task, audio distraction and time pressure. The secondary task and the time pressure both led to a significant reduction in the number of syllables per utterance. Perhaps unsurprisingly, articulation rate was higher, pauses were less frequent and latency was shorter under time pressure. Interestingly though, frequencies of hesitations and disfluencies were significantly higher when performing a secondary task. The audio distraction did not affect significantly any of the symptoms evaluated.

Following the efforts on studying symptoms while subjects had to perform a task on their own, the ComParE paralinguistic challenge implemented a cognitive load automatic detection task in 2014 [30]. The data from this challenge includes samples from different subtasks specifically developed to induce cognitive load: a reading task with ungrammatical content and two variants of the Stroop task: one with time pressure and one involving a secondary task. Using this dataset, [31] found significantly higher values for F0 and for intensity, and significantly lower values for silence durations in the time pressure Stroop condition. When predicting the condition, [31] achieved better results when the articulation rate was used instead of prosodic features (F0, intensity and silence durations), despite not reporting significant differences between conditions for articulation rate.

2.2 Multi-Party Human-Human (MP H-H)

Studying cognitive load symptoms in multi-party human-human settings has been less explored. In [16], we find a study where language from meetings of a fire rescue teams is analysed. High levels of cognitive load were induced through the complexity and the level of priority of the mission that the emergency team was trying to accomplish. Although the interactions here are between humans, this scenario shares some similarities to what we expect to see in ORCA. The observed symptoms include: number of words per utterance, the number of agreement and disagreement expressions and the use of personal pronouns. Agreement expressions were found to be more common in low-load tasks, as well as longer utterances. In addition, variations were observed in the way people use personal pronouns in the different cognitive load conditions.

2.3 Human-Human vs Human-Computer (H-H vs H-C)

Symptoms of cognitive load have also been investigated for human-human and human-computer interactive scenarios. In [3], an in-car setup is used where subjects talk either with the passenger or with a dialogue system. Interactions with a dialogue system were considered to induce a higher cognitive load. A significantly higher value F0 was found when interacting with the dialogue system. Values for the centre frequencies in voiced segments for formants 1 and 4 were also significantly higher in the same condition, together with other spectral measures and the durations of voiced segments.

3 DATA

The data used in this study was introduced in [8] and aimed at studying the effect of cogntive load in Human-Computer interaction scenario. During the collection, the subject is a driver simultaneously interacting with a spoken dialogue system, which provides restaurant information in Cambridge, UK. Although this scenario might not share many similarities with our target scenarion, it is to the best of our the knowledge the only data aimed at studying the effect of cognitive load in H-C interaction.

A within-subjects set-up was used, where driving and speaking was deemed to have higher cognitive load than speaking to the system in isolation. The subjects had to complete 14 predefined requests for different restaurants in Cambridge (7 in each condition). 28 subjects took part in the data collection, which resulted in 390 dialogues. The performance of the subjects on the task of finding an appropriate restaurant was compared but no significant differences were found between conditions.

4 LINGUISTIC SYMPTOMS

We summarise the features from the background works in Table 1, dividing the studies with respect to the type of set-up used in the data collection (columns of Table 1), and discuss them here below in terms of prosodic, turn-taking and linguistic features (rows of Table 1). We then derive results for these features using the dataset from [8] as a test case corpus, some of which have already been reported by the authors. We then explore new symptoms, as described in Section 4.1, using the same corpus.

Prosody. Inspired by previous work [31], we have computed pitch (F0) and intensity (in dB) using Praat [2]. We computed the functionals (mean, standard deviation, skewness, kurtosis, minimum and maximum) for each utterance. In addition, in a similar way to what was done in [31], we plotted the values over time of pitch and intensity. We found the best fit line for these points and took the slope of this line and the mean square error (MSE) between the line and the actual values. In addition, for each turn we have also computed

articulation rate, number of pauses, number of syllables and speech rate duration using the method described in [5]. Higher intensity in the high cognitive load condition was a symptom identified in [8].

Turn-taking. The turn-taking feature group includes silence durations, latency and barge-ins. To compute the silence features, we have first performed Voice Activity Detection for every user turn and we computed the functionals of the duration of the silence segments in an utterance, slope and MSE in the same way we did for F0 and intensity. The latency for each turn is the difference between the time the system stopped speaking and the time the system started speaking. Negative latencies mean that the user started speaking before the system finished the answer and therefore we consider these as barge-ins. In [8], the frequency of barge-ins was higher in the higher load condition.

Linguistic features. Linguistic features often associated with high cognitive load are: disfluencies [14, 21], filler words [8] and filled pauses [14]. However, these are often subjective and difficult to automatically detect with the current state-of-the-art methods. Therefore, the only feature that we will rely on from those previously investigated in the literature is the utterance length in number of words. This was found as a possible indicator for cognitive load in [16], therefore, we extracted it both for the output of the automatic speech recognition (ASR) and the transcription (in order to assess the impact of speech recognition errors in the symptoms). [8] also found differences in the number of turns which followed the task.

4.1 Novel Symptoms

In addition to the features described in Section 4, we discuss here a number of linguistic features, which to the best of our knowledge, have not been previously investigated as symptoms of cognitive load. Some of these refer to the consequences of behaviour symptoms that may not be overtly evident in the speech signal but may affect the spoken dialogue system.

Firstly, the ASR confidence score was extracted from the system logs. The hypothesis is that under high cognitive load, users' speech is going to be harder to recognise and thus will have lower confidence scores, due to the acoustic features mentioned above such as higher articulation/speech rate, higher F0 and increased intensity. For instance, variability of acoustic parameters in children's speech can negatively impact from two to five times the performance of a normal speech recogniser [27]. To further explore this hypothesis, we have also computed the Word Error Rate (WER), in order to evaluate the reliability of the confidence scores.

We secondly hypothesise that higher cognitive load may affect the quality of the interaction. To investigate this, we look at patterns observed in the dialogue acts, shown previously to be indicative of problematic dialogue when interactive with a Spoken Dialogue System [12, 25]. Specifically, we computed the number of slots per user turn and the frequency of both system and user dialogue acts over the whole dialogue.

Finally, we computed the entropy of user/system utterances. Entropy is associated with the amount of information in an utterance. The hypothesis is that if users have higher cognitive load they may choose utterances that have lower information content. The converse could also be true, in that with higher cognitive load the users drop the effort to simplify their utterances with the system (a common phenomena when talking to Spoken Dialogue Systems) in an attempt just to get the information across. We also investigate the system utterances and the difference between adjacent pairs of utterances to see if there is a relation between complexity of system and user utterances. In order to compute the entropy of an utterance, we have used a different dataset but in the same domain [33] to train two different trigram language models, one with user utterances and a second one with the system prompts. With these language models, we compute the likelihood of each utterance given and then the entropy values.

5 RESULTS

In order to provide a meaningful comparison between the feature values in the two conditions, all features were computed at the level indicated in Table 1 and normalised per speaker. To verify if there was a significant difference between conditions (high/low load), we have tested for significance using a t-test when feature values were continuous, Chi-Square test for the frequency features and binomial test for binary features (e.g. barge-in or not in a turn).

Features used in [8]. We found a significantly higher value for intensity in the high cognitive load condition ($M = -0.05, SD = 0.30$) than in the low cognitive load condition ($M = -0.08, SD = 0.29$), $t(2740) = 2.74, p < 0.001$. The number of barge-ins was also significantly higher in the high cognitive load condition .57 than .5 (barge-ins would occur equally in both conditions), $p \approx 0$ (2-sided).

Features previously investigated in the literature. We found significantly higher F0 mean value for turns in the high cognitive load condition ($M \approx 0, SD = 0.47$) when compared with the low cognitive load condition ($M = -0.05, SD = 0.48$), $t(2740) = 2.64, p < 0.01$ and a significantly higher F0 maximum value also in the high load condition ($M = 1.89, SD = 0.34$), when compared with the low cognitive load condition ($M = 1.83, SD = 0.33$), $t(2740) = 5.2, p \approx 0$. For the remaining prosodic features, the tests performed have not revealed any significant differences, however, the trends were that utterances produced in the high cognitive load condition were longer (in number of syllables) and slower (lower rates). With regards the turn taking features, we found silence duration skewness to be significantly higher in the low cognitive load condition ($M = 1.65, SD = 0.54$) when compared with the high cognitive load condition ($M = 1.53, SD = 0.70$), $t(2688) = 2.3, p < 0.05$.

Novel features. We found that the entropy of system prompts in the high load condition was significantly higher ($M = -0.10, SD = 0.13$) when compared the low cognitive load condition ($M = -0.12$, $SD = 0.14$), $t(3168) = 2.5, p < 0.05$. Additionally, we also found that the frequency of the dialogue act *request* in user turns in the low cognitive load condition was significantly higher than in the high cognitive load condition $\chi^2(8, N = 389) = 19.25, p < 0.05$. This means that there were more user utterances where the users were asking for items such as phone number, postcode, address or area after finding a relevant restaurant.

In the first exchange, the system asks the user "How can I help you?". The formulation of the user response may require a higher cognitive effort from the user as it is an open question requiring planning on the user's part. Thus, we analysed the turn level features for this initial user turn in isolation. We found significantly higher skewness for intensity in the low cognitive load

Category	Symptom	Computation Level	H	H-H vs H-C	MP H-H	H-C
Prosodic	Increased intensity	Turn (functionals, slope and MSE)	[29, 31]			[8], TS
	Increased F0 (*)	Turn (functionals, slope and MSE)	[29, 31]	[3]		TS
	Decreased Articulation Rate (*)	Turn (mean)	[14, 31]			
	Increased frequency of Pauses (*)	Turn	[14]		[16]	
	Decreased Number of Syllables (*)	Turn	[14]			
	Decreased Speech Rate (*)	Turn (mean)	[29]			
Turn Taking	Increased Silence Duration (*)	Turn (functionals, slope and MSE)	[17, 31]			TS
	Decreased Latency (*)	Turn	[14, 17]			
	More Frequent Barge-ins	Turn				[8], TS
Linguistic	Utterance Length (*)	Turn (ASR and transcription)			[16]	TS[1]
	Decreased ASR Confidence	Turn				
	Decreased WER	Turn				
	Number of Slots	Turn (System and User)				
	Dialogue Acts	Dialogue (frequency)				TS[2]
	Decreased Entropy	Turn (system, user and difference)				TS[3]
	Fillers(+)	Average per Speaker				[8]
	Following the Task(+)	Turn-by-turn				[8]

Table 1: List of features organised by category with the level which they were extracted and studies they have been found as symptoms for high cognitive load. (*) indicates features from the literature applied to dataset in [8]. (+) indicates features used in [8] but not in this study. Boldface indicates novel features. TS = This study.

condition ($M \approx 0, SD = 0.36$), when compared with the high cognitive load condition ($M = 0.08, SD = 0.39$), $t(391) = 2.24, p < 0.05$ and a significantly higher number of words per utterance in the high cognitive load condition ($M = 1.43, SD = 1.02$) when compared with the low cognitive load condition ($M = 1.21, SD = 1$), $t(391) = 2.14, p < 0.05$, both using the most likely ASR result and transcription. In the following section, we will discuss the implications of these results.

6 DISCUSSION AND FUTURE WORK

In this paper, we have investigated symptoms that could potentially help track the user cognitive load on-the-fly. We have used the corpus described in [8] and extended this work in terms of features discussed in the literature and new features. We have confirmed previous findings reported in [8] for significant different values in intensity and barge-ins. We have found further significantly differences with respect to features in previous studies including F0, utterance length and silence duration, as well as, new features of entropy of system utterances and number of user *requests*.

The differences in system utterance entropy and number of user *requests* could be explained by the fact that there is a higher percentage of utterances in line with the task for subject in the low cognitive load condition as found in [8]. Users behaved as expected, being more predictable and fulfilling the task requirements.

Interesting results were found for the more open "How may I help you?" prompt in terms of utterance length. This leads us to believe that more mixed-initiative dialogues, rather than system-initiative, may result in the user having to plan more and thus induce a higher cognitive load, which in turn might affect symptoms that are observable. If these open questions occur early on in the dialogue, as with the Cambridge dataset, this opens up the possibility for the system to adapt early on in the conversation to the user's cognitive state. Interestingly, this finding contradicts what was previously found in [16], where utterances in the higher cognitive load scenarios were shorter. A possible explanation might be connected with the scenario (fire rescue), which that case, is deemed to be very stressful and therefore people quickly adapt in order to make communication more efficient by making their utterances shorter. In [14], authors have also found different tendencies depending on the way they induced cognitive load. This seems to indicate that the symptoms are highly dependent both on the nature of the main task and the way the cognitive load is induced.

Designing a task that is engaging and varied enough not to entrain the user (and thus naturally reduce the load) will be a challenge that we will have to address in our future data collections. Performance metrics were not discussed in detail here and previous literature does not report significant results in this regard. However, performance of the user and the system is key for the scenario for the ORCA project and therefore will be re-examined in future work. In addition, the effect of cognitive load during task-oriented vs open domain (chitchat dialogue) would be an interesting research question to explore in future work but would be outside the ORCA project domain, which is very much task-orientated (e.g. performing inspections, following procedures, situation awareness). As the scenarios that will be implemented will potentially involve emergencies, we expect that differences in prosodic features to be more visible when the situation becomes critical, than the differences observed in the data analysed in this study. Regarding linguistic features we also expect results in line with [16] for scenarios that induce high cognitive load rather than those observed in this study.

In the future, we want to continue exploring features that were found in other studies such as spectral features [19], disfluencies and filled pauses [14, 21]. Given that the multi-modal nature of the interactions in the scope of the ORCA project, we could aim to explore other features reported in the literature such as pupil dilation [13] as we believe that this would provide further information with regards the user state to which the system could adapt.

[1] After the "How may I help you?" prompt.
[2] For user *request* dialogue act.
[3] For system prompts.

ACKNOWLEDGMENTS

We would like to thank Dr. Milica Gašić and the authors of [8] for access to their dataset and assistance. This research was funded by EPSRC ORCA Hub (EP/R026173/1, 2017-2021); RAEng/Leverhulme Trust Senior Research Fellowship Scheme (Hastie/ LTSRF1617/13/37).

REFERENCES

[1] Richard W Backs and Larry C Walrathf. 1992. Eye movement and pupillary response indices of mental workload during visual. *Applied ergonomics* 23, 4 (1992), 243–254.

[2] Paul Boersma and David Weenink. 2018. Praat: doing phonetics by computer [Computer Program]. Version 6.0.39 retrieved in May 2018. http://www.praat.org

[3] Hynek Bořil, Seyed Omid Sadjadi, Tristan Kleinschmidt, and John HL Hansen. 2010. Analysis and detection of cognitive load and frustration in drivers' speech. In *Eleventh Annual Conference of the International Speech Communication Association*.

[4] Fang Chen, Natalie Ruiz, Eric Choi, Julien Epps, M Asif Khawaja, Ronnie Taib, Bo Yin, and Yang Wang. 2012. Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 4 (2012), 22.

[5] Nivja H De Jong and Ton Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods* 41, 2 (2009), 385–390.

[6] Vera Demberg, Asad Sayeed, Angela Mahr, and Christian Müller. 2013. Measuring linguistically-induced cognitive load during driving using the ConTRe task. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 176–183.

[7] Francisco J. Chiyah Garcia, David A. Robb, Atanas Laskov, Xingkun Liu, Pedro Patron, and Helen Hastie. [n. d.]. Explainable Autonomy: A Study of Explanation Styles for Building Clear Mental Models through a Multimodal Interface.

[8] Milica Gašić, Pirros Tsiakoulis, Matthew Henderson, Blaise Thomson, K. Yu, E. Tzirkel, and Steve Young. 2012. The Effect of Cognitive Load on a Statistical Dialogue System. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 74–78.

[9] Helen Hastie, Francisco J. Chiyah Garcia, David A. Robb, Atanas Laskov, and Pedro Patron. 2017. MIRIAM: A Multimodal Chat-based Interface for Autonomous Systems. In *Proc. ICMI'17*. 495.

[10] Helen Hastie, Xingkun Liu, and Pedro Patron. 2017. Trust triggers for multimodal command and control interfaces. In *Proc. ICMI'17*. 261–268.

[11] Helen Hastie, Katrin Lohan, Mike Chantler, David A. Robb, Ron Petrick, David Lane, Subramanian Ramamoorthy, and Sethu Vijayakumar. 2018. The ORCA Hub: Explainable Offshore Robotics through Intelligent Interfaces. *Workshop on Explainable Robotic Systems. HRI Conference* (2018).

[12] Helen Wright Hastie, Rashmi Prasad, and Marilyn Walker. 2002. What's the trouble: automatically identifying problematic dialogues in DARPA communicator dialogue systems. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 384–391.

[13] Shamsi T Iqbal, Piotr D Adamczyk, Xianjun Sam Zheng, and Brian P Bailey. 2005. Towards an index of opportunity: understanding changes in mental workload during task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 311–320.

[14] Anthony Jameson, Juergen Kiefer, Christian Müller, Barbara Großmann-Hutter, Frank Wittig, and Ralf Rummer. 2010. Assessment of a user's time pressure and cognitive load on the basis of features of speech. In *Resource-adaptive cognitive processes*. Springer, 171–204.

[15] David O Kennedy and Andrew B Scholey. 2000. Glucose administration, heart rate and cognitive performance: effects of increasing mental effort. *Psychopharmacology* 149, 1 (2000), 63–71.

[16] M Asif Khawaja, Fang Chen, and Nadine Marcus. 2012. Analysis of collaborative communication for linguistic cues of cognitive load. *Human factors* 54, 4 (2012), 518–529.

[17] M Asif Khawaja, Natalie Ruiz, and Fang Chen. 2007. Potential speech features for cognitive load measurement. In *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces*. ACM, 57–60.

[18] Andrew L Kun, Oskar Palinko, and Ivan Razumenić. 2012. Exploring the effects of size and luminance of visual targets on the pupillary light reflex. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 183–186.

[19] Phu Ngoc Le, Eliathamby Ambikairajah, Julien Epps, Vidhyasaharan Sethu, and Eric HC Choi. 2011. Investigation of spectral centroid features for cognitive load classification. *Speech Communication* 53, 4 (2011), 540–551.

[20] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. 2119–2129. https://doi.org/10.1145/1518701.1519023

[21] Anders Lindström, Jessica Villing, Staffan Larsson, Alexander Seward, Nina Åberg, and Cecilia Holtelius. 2008. The effect of cognitive load on disfluencies during in-vehicle spoken dialogue. In *Ninth Annual Conference of the International Speech Communication Association*.

[22] Ottmar V Lipp and David L Neumann. 2004. Attentional blink reflex modulation in a continuous performance task is modality specific. *Psychophysiology* 41, 3 (2004), 417–425.

[23] Katrin S. Lohan, Eli Sheppard, G. E. Little, and Gnanathusharan Rajendran. 2018. Towards improved child robot interaction by understanding eye movements. *IEEE Transactions on Cognitive and Developmental Systems* (2018), 1–1. https://doi.org/10.1109/TCDS.2018.2838342

[24] Sandra P Marshall, CW Pleydell-Pearce, and BT Dickson. 2002. Integrating psyhophysiological measures of cognitive workload and eye movements to detect strategy shifts. In *HICSS '03 Proceedings of the 36th Annual Hawaii International Conference on System Sciences*. Citeseer.

[25] Raveesh Meena, José Lopes, Gabriel Skantze, and Joakim Gustafson. 2015. Automatic detection of miscommunication in spoken dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 354–363.

[26] Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal WM Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist* 38, 1 (2003), 63–71.

[27] Alexandros Potamianos and Shrikanth Narayanan. 2003. Robust recognition of children's speech. *IEEE Transactions on speech and audio processing* 11, 6 (2003), 603–616.

[28] David A. Robb, Francisco J. Chiyah Garcia, Atanas Laskov, Xingkun Liu, Pedro Patron, and Helen Hastie. 2018. Keep Me in the Loop: Increasing Operator Situation Awareness through a Conversational Multimodal Interface. In *Proc. ICMI'18*. https://doi.org/10.1145/3242969.3242974

[29] Klaus R Scherer, Didier Grandjean, Tom Johnstone, Gudrun Klasmeyer, and Thomas Bänziger. 2002. Acoustic correlates of task load and stress. In *Seventh International Conference on Spoken Language Processing*.

[30] Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. 2014. The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[31] Maarten Van Segbroeck, Ruchir Travadi, Colin Vaz, Jangwon Kim, Matthew P Black, Alexandros Potamianos, and Shrikanth S Narayanan. 2014. Classification of cognitive load from speech using an i-vector framework. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[32] Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 extended abstracts on Human factors in computing systems*. ACM, 2651–2656.

[33] T. Wen, N. Mrksic, and S. Young. 2016. Research data supporting "Conditional Generation and Snapshot Learning in Neural Dialogue Systems" [Dataset]. https://doi.org/10.17863/CAM.6142

[34] Glenn F Wilson and Christopher A Russell. 2003. Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human factors* 45, 4 (2003), 635–644.

[35] R. H. Wortham, A. Theodorou, and J. J. Bryson. 2017. Improving robot transparency: Real-time visualisation of robot AI substantially improves understanding in naive observers. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 1424–1431. https://doi.org/10.1109/ROMAN.2017.8172491