Combining Oculo-motor Indices to Measure Cognitive Load of Synthetic Speech in Noisy Listening Conditions.

MATEUSZ DUBIEL*, The University of Strathclyde
MINORU NAKAYAMA, Tokyo Institute of Technology
XIN WANG, National Institute of Informatics

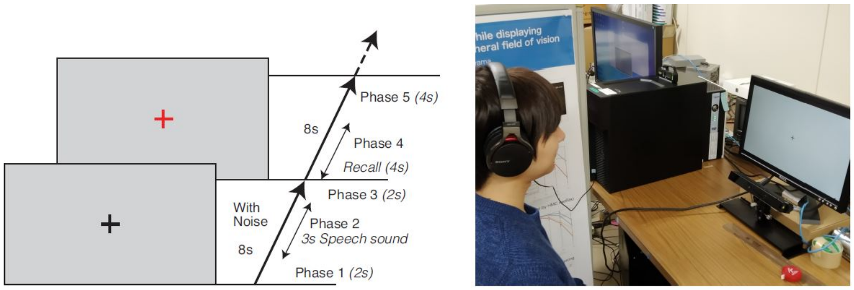

Fig. 1. An overview of the experimental procedure (left) and illustration of the experimental setup featuring one of the participants during the recall task (right).

Gaze-based assistive technologies (ATs) that feature speech have the potential to improve the life of people with communication disorders. However, due to a limited understanding of how different speech types affect the cognitive load of users, an evaluation of ATs remains a challenge. Expanding on previous work, we combined temporal changes in pupil size and ocular movements (saccades and fixation differentials) to evaluate cognitive workload of two types of speech (natural and synthetic) mixed with noise, through a listening test. While observed pupil sizes were significantly larger at lower signal-to-noise levels, as participants listened and memorised speech stimuli; saccadic eye-movements were significantly more frequent for synthetic speech. In the synthetic condition, there was a strong negative correlation between pupil dilation and fixation differentials, indicating a higher strain on participants' cognitive resources. These results suggest that combining oculo-motor indices can aid our understanding of the cognitive implications of different speech types.

CCS Concepts: • **Human-centered computing** → *Laboratory experiments*; *Empirical studies in HCI*.

Additional Key Words and Phrases: Speech Perception, Pupil response, Eye movement, Evaluation study

ACM Reference Format:
Mateusz Dubiel, Minoru Nakayama, and Xin Wang. 2021. Combining Oculo-motor Indices to Measure Cognitive Load of Synthetic Speech in Noisy Listening Conditions.. In *ETRA '21: 2021 Symposium on Eye Tracking Research and Applications (ETRA '21 Short Papers), May 25–27, 2021, Virtual Event, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3448018.3458013

*corresponding author

1 INTRODUCTION

Gaze-based assistive technologies (ATs) provide opportunity to support independence of people with motor and learning disabilities. Examples of ATs devices that feature synthetic speech include gaze-triggered reading assistance interfaces that facilitate vocabulary learning [Sibert et al. 2000] or Augmentative and Assisted Communication (AAC) devices that enable spoken communication [Caligari et al. 2013; Kane et al. 2017]. Despite strong potential of ATs to support everyday life activities, research shows that some user groups may experience challenges in understanding synthetic speech [Hux et al. 2017]. Adoption of AAC devices is currently hindered by poor interface design and lack of usability testing [Ascari et al. 2020]. While obtaining accurate insights of users cognitive workload is crucial for improving interface usability and increasing the likelihood of its adoption, identifying objective and non-intrusive metrics is a challenging task [Fridman et al. 2018]. Eye technologies offer an opportunity to address this challenge by using various characteristics of the ocular reactivity and the eye movement in response to varying cognitive workload.

Our work presents how a combination of oculo-motor indices can be used to evaluate cognitive workload of Speech-In-Noise (SIN) listening tasks for natural and synthetic Japanese speech. We hope that findings of our investigation will be valuable to the COGAIN community and help to foster development of robust AT evaluation metrics in the future.

2 BACKGROUND

Cognitive Load Theory (CLT) is based on the premise that cognitive capacity in working memory is limited [Sweller 1988]. In ergonomics, the concept of cognitive workload is used to describe internal processing of task that cannot be observed directly [Kruger et al. 2013]. CLT plays a crucial role in design of interactive interfaces as it can provide insights on how to avoid overloading users [Duchowski et al. 2018]. In this section, we provide an overview of eye-tracking metrics used for estimation of cognitive workload and discuss their application to evaluate synthetic speech systems.

2.1 Eye-related Cognitive Workload Metrics

Using pupilometry to measure cognitive workload has a long research tradition. Early experiments showed that increasing task difficulty in manipulation tasks leads to increasing pupil diameter (pupil dilation) [Hess and Polt 1964]. Later studies indicated that overly difficult tasks led to information overload but did not result in pupil dilation [Peavler 1974]. Following several decades of research, task-evoked pupillary dilation was accepted as a reliable and sensitive index of within-task variations in processing load [Beatty 1982]. More recently, using frequency of pupil diameter oscillation has been presented as an effective alternative to pupil dilation metric [Duchowski et al. 2020, 2018]. The advantage of the pupil oscillation approach is that, unlike the pupil dilation approach, it is not a baseline-related measure which could be affected by pupil sensitivity to different illumination levels. Another problem is that pupil diameter, as measured by the eye-tracker, undergoes variations during movement of the eye which could impact the evaluation.

Besides pupil-based methods, other eye-tracking approaches to cognitive workload evaluation focus on positional eye movements, i.e. motor indices. These include: number and frequency of fixations (movements of eye when focusing on a target) [Jacob and Karn 2003], duration of fixations [Just and Carpenter 1976], and microsaccades (miniature eye movements that occur involuntarily during fixation) [Engbert and Kliegl 2003]. In current work, we combine pupillary responses with motor indices to provide insights into cognitive load during a 'listen and recall' SIN experiment.

2.2 Evaluating Synthetic Speech

Pupillometry has recently been applied to evaluation of text-to-speech (TTS) systems [Govender and King 2018a,b; Govender et al. 2019a; Simantiraki et al. 2018]. Experiments on English speech indicated that in quiet listening conditions increased pupil dilation indicates attention and engagement [Govender and King 2018a], while in noisy listening conditions, increased pupil dilation indicates increased listening effort [Govender et al. 2019b]. Results of the study that explored the impact of speech enhancement algorithms in ‘listen and recall’ task, showed that enhanced speech can reduce cognitive workload and increase correctness of recall accuracy [Simantiraki et al. 2018]. Different types of speech synthesisers from low quality Hidden Markov systems to high quality Deep Neural Network systems were found to evoke higher cognitive workload than natural speech under noisy listening conditions [Govender et al. 2019a].

We applied pupillometric evaluation approach to measure differences in cognitive workload between natural and synthetic Japanese speech mixed with noise (see [Dubiel et al. 2020, 2021] for detailed analysis of results). Pupil dilations and oscillations observed during our experiment revealed significant differences between two experimental conditions: (1) type of speech (natural vs. synthetic) and (2) level of signal-to-noise (-1dB, -3dB, -5dB). We postulate that the observed pupillary responses may also be linked to other oculo-motor indices. In this paper, we analyse saccade occurrences, saccade frequencies and fixation differentials observed during the experiment, and explore their relationship to pupilary changes. Our investigation provides new insights into cognitive implications of different types of speech during a SIN ‘listen and recall’ task that can contribute to more robust usability testing of gaze-based interfaces that feature speech.

3 EXPERIMENT

The goal of the experiment was to evaluate the impact of type of speech (natural vs. synthetic) and signal-to-noise level (-1dB, -3dB, -5dB) on participant’s cognitive workload in a ‘listen and recall’ task. In the evaluation, we used a combination of subjective (questionnaires) and objective (oculo-motor responses) metrics. In the current paper, we focus on the analysis of objective metrics and discussion of the relationship between pupil diameter-based and eye movement-based indices. Please see [Dubiel et al. 2021] for detailed discussion of the results obtained from subjective metrics. In design of our experiment, we followed recommendations provided in [Winn et al. 2018].

3.1 Speech Stimuli

40 travel-related speech samples were selected from a Japanese speech corpus of Saruwatari-lab., University of Tokyo (JSUT) [1]. The selected speech samples were sampled at 48kHz, synthesised using a state-of-the art Japanese TTS system, and mixed with speech-shaped noise at three levels of signal-to-noise (-1dB, -3dB, -5dB). The TTS system was built on the basis of the classical neural-network-based statistical parametric speech synthesis framework [Zen et al. 2013]. It uses an OpenTalk-based text analyzer [HTS Working Group 2015] to convert the text string into phonetic labels, a Recursive Neural Network acoustic model to convert the labels into acoustic features (i.e., Mel-cepstral coefficients, fundamental frequency trajectory, and aperiodicity parameters), and the WORLD vocoder [HTS Working Group 2015] to produce the 48 kHz waveform with acoustic features. Following an informal listening test by the authors, 10 samples were selected from the synthesised stimuli to ensure high quality (i.e. no unnatural pauses and no mispronunciations). Finally, 10 natural speech samples were selected based on the syntactic similarity to the 10 selected synthetic samples.

[1] https://sites.google.com/site/shinnosuketakamichi/publication/jsut

3.2 Procedure

The experimental procedure is presented in Figure 1. Participants were assigned into one of three signal-to-noise levels and listened to two blocks of speech stimuli (natural and synthetic), presented in a counterbalanced order. Each bloc contained 10 samples, played in a fixed order. The experimental procedure consisted of five consecutive phases. Participants were asked to look at a black fixation cross on a grey background, listen to the speech stimulus (phase 2, 3s), retain it in their memory (phase 3, 2s), and repeat what they heard when the cross changed colour to red (phase 4, 4s). Finally, phase 5 (4s) was time to 'relax and refresh' before the next stimulus. Masking speech-shaped-noise was present in phases 1 to 3. The recall attempt was considered successful only if the whole utterance was repeated correctly.

Participants' eye movements were measured using an eye-tracker (nac:ACTUS, sampling rate = 60Hz). The tracked data for both eyes were processed as a repeated measure for each participant. During the experiment the 'relative pupil size' was calculated by dividing the 'the observed pupil size' by the 'baseline'. The baseline was measured during phase 1 of the experiment. In addition to the temporal pupillary changes, metrics of eye movements (i.e. saccade occurrence and saccade frequency) were extracted using a threshold $40deg/sec$. (as recommended by [Ebisawa and Sugiura 1998]). We also measured shifts in participants' fixation points while they were looking at the fixation cross (+). The experimental setup is presented in Figure 1. The room illumination and screen brightness were kept constant for every trial.

3.3 Subjects

16 native Japanese speakers (M = 14, F = 2) with no self-reported hearing or vision problems took part in the experiment. The participants were aged between 21 and 25 (mean = 22.5 years). The participants were split into 3 groups (-1dB = 6, -3dB = 5 and -5dB = 5). All participants attended a briefing session before the experiment and provided their written consent in order to participate. Each participant was given 1000 yen for taking part in the experiment.

4 RESULTS

Table 1. Comparison of Recall Accuracy. '**' signifies p < 0.01

Signal-to-Noise Level	Speech Type	Mean	STD
-1dB**	Natural	1.00	-
	Synthetic	0.74	0.11
-3dB	Natural	0.96	0.09
	Synthetic	0.78	0.16
-5dB**	Natural	0.90	0.10
	Synthetic	0.68	0.05

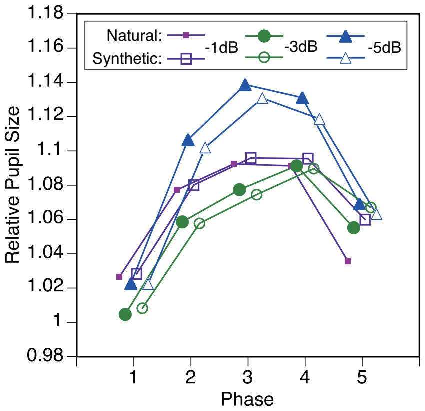

Fig. 2. Pupillary changes for 5 experimental phases.

4.1 Recall accuracy

Participants' performance in recall of speech stimuli is presented in Table 1. The performance was evaluated by one of the experimenters who listened to participants' responses and classified them into 'correct' and 'incorrect' categories. There was a significant difference in recall rates between natural and synthetic except for the -3dB condition (-1dB: t(4)

= 5.1, p < 0.01, -3dB : t(8) = 2.2, p = 0.06, -5dB : t(7) = 4.1, p < 0.01). Two-way ANOVA indicated that the factor of speech type is significant (F(1, 23) = 34.0, p < 0.01) while the factor of speech sound level was not significant (F(2, 23) = 2.01, p = 0.16). These results indicate that synthetic speech puts more strain on cognitive resources of the participants. In the following subsections we provide analysis of oculo-motor indices based exclusively on correct the recall responses.

4.2 Pupillary responses

The pupil sizes for both experimental conditions (i.e. speech type and signal level), across all 5 phases are summarised in Figure 2. The figure features the correct recall attempts, i.e. attempts where participants managed to repeat whole spoken stimulus without making any mistakes. As can be seen in Figure 2, pupil size increases until phase 3 and decreases from thereon until phase 5. Since pupil responses have a 100-200ms delay [Beatty 1982], the size for phase 4 is also the highest in some conditions. At phase 5 (after the recall phase), the pupil sizes for synthetic condition are higher than for natural condition with the exception of -5dB signal-to-noise level. The fact that at -5dB pupil dilation for natural speech was higher than for synthetic speech could be attributed to participants exceeding their 'attention ceiling' - this phenomenon was reported in [Simantiraki et al. 2018].

Our statistical analysis focused on phase 5, where the biggest differences in pupil dilations between natural and synthetic speech were detected. We conducted a two-way ANOVA to determine the impact of speech type (natural vs. synthetic) and three signal-to-noise levels (-1 ~ -5dB) on pupil dilation. Both factors were found to significantly affect pupil size (speech type: $F(1, 348) = 5.91, p < 0.05$; sound level: $F(2, 348) = 3.14, p < 0.05$). The interaction was not significant. Mean pupil size for Synthetic speech was larger than for Natural speech, and also pupil size was shown to increase with the decrease in level of signal (-1dB to -5dB). As participants recalled the presented stimuli (phase 3), their mental effort was manifested in increased pupil size. There were no significant differences between natural and synthetic speech in other experimental phases.

We also compared frequency spectrum densities of pupillary oscillations across phases for both conditions. We found significant differences ($p < 0.05$) between natural and synthetic speech at -3dB and and -5dB levels. However, the orders were different between the conditions; at -3dB, oscillations were higher for natural than synthetic speech and for -5dB the reverse was the case. This result may indicate that participants' attention ceiling was reached at the lowest signal-to-noise condition. A more detailed discussion of pupil oscillations is presented in [Dubiel et al. 2021].

4.3 Eye movements analysis

Participants' eye movements were classified into fixations and saccades using a 40deg/sec. threshold [Ebisawa and Sugiura 1998]. Features of both saccades and fixations were summarised for every experimental phase. In this section, we present the analysis of recorded saccades and fixations, and compare them across both experimental conditions.

4.3.1 Saccade occurence. The overall occurrence rates of saccades are presented in Figure 3. Saccade occurrence rate corresponds to the proportion of a phase that contains saccades. Since our experimental task requires attention, saccades should be suppressed during the memorising phrase [May et al. 1990]. Conversely, the occurrence of saccades may indicate a break of attention. As shown in Figure 3, saccade occurrences decrease monotonically from phase 1 to phase 3, while subjects memorise and retain the speech stimuli. After phase 3, saccade occurrences increase rapidly and stay high towards the end of a trial. Since the retention of information puts strain on cognitive resources, saccade eye movements are suppressed. During phase 3 there are some differences between the experimental conditions. We tested the impact of both types of speech on saccade occurrence (1/0) using χ^2 test. There were significant differences

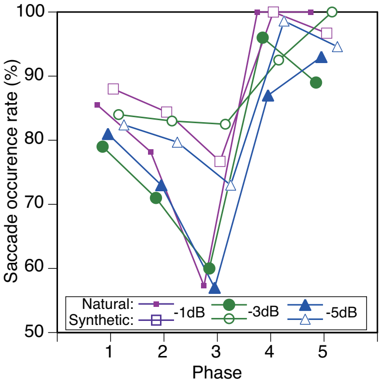

Fig. 3. Rates of saccade occurrence across experimental conditions.

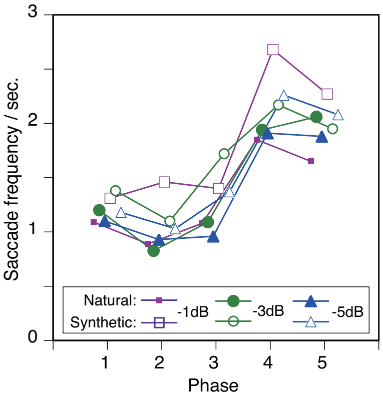

Fig. 4. Saccade frequencies in response to the experimental conditions.

between natural and synthetic speech in phase 3 (-1dB:χ^2=8.9,$p < 0.01$; -3dB:χ^2=12.3,$p < 0.01$; -5dB:χ^2=4.3,$p < 0.05$). This indicates that synthetic speech may be a trigger for saccades. However, there were no statistically significant differences between signal-to-noise levels.

5 RESULTS

5.0.1 Saccade frequencies. Saccade frequencies for every phase were calculated for both experimental conditions. Here, saccade frequency is considered as a number of saccades per second. High saccade frequency may indicate a decrease in participant's attention due to high cognitive load. The mean saccade frequencies are summarised in Figure 4. The mean frequency of observed saccades was less than 3 which implies that cognitive load evoked by the task was not very high. The mean frequencies were suppressed during phases 2 and 3 as participants focused their attention on memorising speech sounds. A substantial increase of saccade occurrences (Figures 3) and saccade frequencies (Figure 4) at phase-4 indicates that participants released their attention from the task.

The comparison of saccade frequencies between natural and synthetic speech revealed significant differences in all phases except phase 1 (phase 1:$F(1, 534) = 3.22, p < 0.10$; phase 2:$F(1, 534) = 16.0, p < 0.01$; phase 3:$F(1, 534) = 16.2, p < 0.01$; phase 4:$F(1, 534) = 15.89, p < 0.01$; phase 5:$F(1, 534) = 4.1, p < 0.05$). There was also a significant difference between sound levels in phase 2 ($F(2, 534) = 3.25, p < 0.05$). This suggests that both speech types differently influenced saccade eye movements during the 'listen & recall' task.

5.0.2 Fixation differentials. As explained in Section 3.2, during the experiment, participants were asked to focus their gaze on the fixation cross (+). We analysed shifts and drifts of participant's eye during fixation (cf. [Engbert and Kliegl 2003]). Here, we summarise these changes in eye movement as fixation differentials with regards to the sampling rate of the eye tracker used in this experiment. The fixation differentials are summarised for each experimental condition and presented in Figure 5. Although there were no statistically significant differences between the experimental conditions, we observed that fixation differentials were suppressed while speech stimuli are listened to and retained (phase 2 and phase 3). As can be seen, the mean differentials are restricted around the 5 degree mark, with higher deviation for synthetic speech as compared to natural speech.

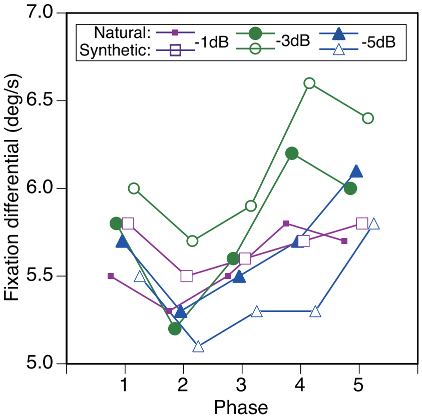

Fig. 5. Fixation differentials in visual angles (*degree*/*second*).

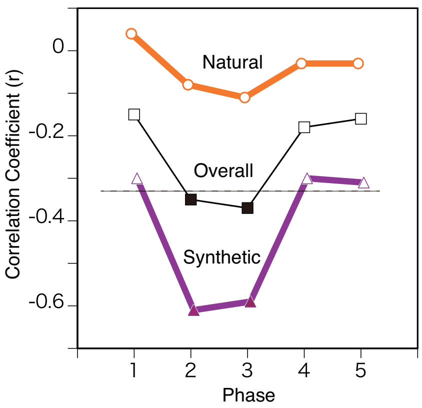

Fig. 6. Correlation coefficients between pupil size and fixation differentials (r).

5.1 Relationship between pupillary responses and eye movements

Observed pupillary responses and eye movements were compared between the experimental conditions, across all phases. In this section, we examine the relationship between the applied evaluation metrics and comment on insights that they provide into participants' cognitive workload experienced during the task.

We observed an exceptional relationship between pupil size and fixation points differentials in the synthetic speech condition. Changes of correlation coefficients for both types of speech are summarised in Figure 6. The vertical axis indicates correlation coefficient (r), and the dotted line indicates level of significance ($p = 0.05$). Overall correlation coefficients are calculated across five phases. There are significant coefficients for phase 2 (r = -0.35, p = 0.048) and phase 3 (r = -0.37, p = 0.04). The correlation coefficients are compared between two conditions: natural and synthetic speech. For synthetic speech, statistically significant negative coefficients appear in phase 2 (r = -0.61, p = 0.01) and phase 3 (r = -0.59, p = 0.02). We observed that while pupil dilates in phases 2 and 3 (see Figure 2) the fixation points were suppressed (see Figure 5). Since in the synthetic condition we observed less suppression of fixation points, it may indicate that this condition was more challenging and required participants to exert more cognitive effort.

6 DISCUSSION AND CONCLUSION

This paper investigated the use of different oculo-motor indices as metrics of cognitive workload in a SIN 'listen and recall' experiment. Our study expanded the pupillometric approach to evaluation of synthetic speech by also analysing other eye-movement metrics and exploring their relationship with pupil dilation. While pupil responses were found to be sensitive to the difference in signal-to-noise level, eye movements had a distinctive behaviour for each speech type (natural and synthetic). We demonstrated that eye saccades and fixation differentials can provide additional insights regarding the impact of different types of speech stimuli on participants' cognitive workload and recall accuracy.

We observed that the percentage of saccades that occurred during listening and retention phases was significantly higher for synthetic speech than for natural speech, which suggests that the synthetic condition made it more difficult for participants to focus. A similar trend was found for the frequency of saccades, which was higher for synthetic speech than for natural speech. An interesting finding is that there was a significant strong negative correlation between pupil dilation (an established evaluation metric) and fixation differentials (proposed extension to the evaluation approach) during phase 2 and 3. This indicates that while listening to synthetic speech stimuli and committing them to memory, it was more difficult for participants to focus their gaze on the fixation cross (+). We postulate that using natural speech recordings may be more beneficial for tasks that require keeping sustained focus and attention.

6.1 Implications for AT evaluation

A recent usability survey of ATs revealed that while the most common activity for using gaze-based interfaces was to talk (66%, 111 of 169 participants), 42% of participants (71 of 169) reported using assistive technology to be effortful [Hemmingsson and Borgestig 2020]. The detrimental impact that high cognitive load could have on users was demonstrated in a gaze-typing context where it led to lower typing accuracy [Bafna et al. 2020]. However, usability experiments of gaze-based interfaces are still scarce [Ascari et al. 2020]. Since AAC devices that feature speech have strong potential to increase social involvement of individuals with physical disabilities and complex information needs [McNaughton et al. 2019], it is important to understand the cognitive implications that such devices have on users.

We believe that an evaluation approach that combines different metrics can contribute to improving usability testing of gaze-based ATs by providing a non-intrusive way to measure cognitive workload of interfaces that feature speech. Examples of potential evaluation tasks include: gaze-triggered auditory prompting (e.g. [Sibert et al. 2000]) where either synthetic or natural voice is used to facilitate vocabulary learning; using AAC devices in a noisy setting (e.g. large family gatherings, communicating outdoors etc.), or browsing web with audio content (e.g. gaze-triggered audiobooks).

6.2 Limitations and future work

We are mindful that gaze behaviour on the same target can differ based on the task (cf. [Yarbus 2013, Chapter 7]) and that experience of cognitive workload can vary between individuals [Guastello et al. 2015]. Therefore, it would be important to validate our findings through in-situ studies involving different groups of AAC users. It is also important to note that while recent TTS systems are considered to be almost indistinguishable from human speech [Luong et al. 2019], many AAC devices still offer the outdated HMM speech synthesisers [Tokuda et al. 2002] that are characterised by unnatural prosody and 'robotic' pitch. Therefore, evaluation on interfaces with different TTS system can provide service providers with better understanding of how type of voice can affect cognitive workload and user experience.

In the future, machine learning methods could be used to monitor eye-movement behaviour of users and adapt the volume of speech and level of voice expressivity based on the observed cognitive workload, while taking into account current acoustic conditions in which communication takes place. Examples of possible speech adjustments include using 'Lombard speech' effect [Valentini-Botinhao et al. 2013] or applying a Near End Listening Enhancement algorithm for improved comprehension [Chermaz et al. 2019].

Finally, we would like to point out that the low sampling rate of our eye-tracker (60 Hz) did not allow for detection microsaccades and calculation of saccade speeds. In the future, we plan to conduct further experiments using more accurate measuring instruments with higher sampling rates. We hope that an analysis of microsaccades will provide us with further useful insights for using oculo-motor indices for measuring the cognitive workload of synthetic speech.

6.3 Closing Remarks

Combining oculo-motor indices provides valuable insights that could contribute to better understanding of the impact of speech on cognitive workload. In order to make evaluation of gaze-based assistive technologies more ecologically valid, the future experiments should feature real-life tasks and, where possible, incorporate users' feedback. Collaboration between eye-tracking, speech-processing and human-computer interaction communities can help to enhance usability testing, foster participatory research and, ultimately, lead to development of more user-friendly AAC interfaces.

REFERENCES

Rúbia EO Schultz Ascari, Roberto Pereira, and Luciano Silva. 2020. Computer Vision-based Methodology to Improve Interaction for People with Motor and Speech Impairment. *ACM Transactions on Accessible Computing* 13, 4 (2020), 1–33.

Tanya Bafna, John Paulin Paulin Hansen, and Per Baekgaard. 2020. Cognitive Load during Eye-typing. In *ACM ETRA*. 1–8.

Jackson Beatty. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin* 91, 2 (1982), 276.

Marco Caligari, Marco Godi, Simone Guglielmetti, Franco Franchignoni, and Antonio Nardone. 2013. Eye tracking communication devices in amyotrophic lateral sclerosis: impact on disability and quality of life. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 14, 7-8 (2013), 546–552.

Carol Chermaz, Cassia Valentini-Botinhao, Henning Schepker, and Simon King. 2019. Evaluating Near End Listening Enhancement Algorithms in Realistic Environments. *Proc. Interspeech 2019* (2019), 1373–1377.

Mateusz Dubiel, Minoru Nakayama, and Xin Wang. 2020. *Using Pupillary Responses to Measure Cognitive Load of Japanese Synthetic Speech mixed with Noise*. Technical Report HIP2020-51. IEICE Technical report. 93–96 pages.

Mateusz Dubiel, Minoru Nakayama, and Xin Wang. 2021. Evaluating Synthetic Speech Workload with Oculo-motor indices: Preliminary Observations for Japanese Speech. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC2021)*, Vol. 4:BIOSIGNALS. INSTICC publishing, Lisbon, 335–342.

Andrew T Duchowski, Krzysztof Krejtz, Nina A Gehrer, Tanya Bafna, and Per Bækgaard. 2020. The Low/High Index of Pupillary Activity. In *Proceedings of the 2020 Conference on Human Factors in Computing Systems*. 1–12.

Andrew T Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Proceedings of the 2018 Conference on Human Factors in Computing Systems*. 1–13.

Yoshinobu Ebisawa and Mitsuhiro Sugiura. 1998. Influences of Target and Fixation Point Conditions on Characteristics of Visually Guided Voluntary Saccade. *The Journal of the Institute of Image Information and Television En gineers* 52, 11 (1998), 1730–1737.

Ralf Engbert and Reinhold Kliegl. 2003. Microsaccades uncover the orientation of covert attention. *Vision research* 43, 9 (2003), 1035–1045.

Lex Fridman, Bryan Reimer, Bruce Mehler, and William T Freeman. 2018. Cognitive load estimation in the wild. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–9.

Avashna Govender and Simon King. 2018a. Measuring the Cognitive Load of Synthetic Speech Using a Dual Task Paradigm.. In *Interspeech*. 2843–2847.

Avashna Govender and Simon King. 2018b. Using pupillometry to measure the cognitive load of synthetic speech. *System* 50 (2018), 100.

Avashna Govender, Cassia Valentini-Botinhao, and Simon King. 2019a. Measuring the contribution to cognitive load of each predicted vocoder speech parameter in dnn-based speech synthesis. In *Speech Synthesis Workshop (SSW)*, Vol. 2019.

Avashna Govender, Anita E Wagner, and Simon King. 2019b. Using Pupil Dilation to Measure Cognitive Load When Listening to Text-to-Speech in Quiet and in Noise.. In *INTERSPEECH*. 1551–1555.

Stephen J Guastello, Anton Shircel, Matthew Malon, and Paul Timm. 2015. Individual differences in the experience of cognitive workload. *Theoretical Issues in Ergonomics Science* 16, 1 (2015), 20–52.

Helena Hemmingsson and Maria Borgestig. 2020. Usability of eye-gaze controlled computers in Sweden: A total population survey. *International journal of environmental research and public health* 17, 5 (2020), 1639.

Eckhard H Hess and James M Polt. 1964. Pupil size in relation to mental activity during simple problem-solving. *Science* 143, 3611 (1964), 1190–1192.

HTS Working Group. 2015. The Japanese TTS System Open JTalk. http://open-jtalk.sourceforge.net/

Karen Hux, Kelly Knollman-Porter, Jessica Brown, and Sarah E Wallace. 2017. Comprehension of synthetic speech and digitized natural speech by adults with aphasia. *Journal of Communication Disorders* 69 (2017), 15–26.

Robert JK Jacob and Keith S Karn. 2003. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye*. Elsevier, 573–605.

Marcel Adam Just and Patricia A Carpenter. 1976. Eye fixations and cognitive processes. *Cognitive psychology* 8, 4 (1976), 441–480.

Shaun K Kane, Meredith Ringel Morris, Ann Paradiso, and Jon Campbell. 2017. " At times avuncular and cantankerous, with the reflexes of a mongoose" Understanding Self-Expression through Augmentative and Alternative Communication Devices. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1166–1179.

Jan-Louis Kruger, Esté Hefer, and Gordon Matthew. 2013. Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts. In *Proceedings of the 2013 Conference on Eye Tracking South Africa*. 62–66.

Hieu-Thi Luong, Xin Wang, Junichi Yamagishi, and Nobuyuki Nishizawa. 2019. Training multi-speaker neural text-to-speech systems using speaker-imbalanced speech corpora. *arXiv preprint arXiv:1904.00771* (2019).

James G May, Robert S Kennedy, Mary C Williams, William P Dunlap, and Julie R Brannan. 1990. Eye movement indices of mental workload. *Acta psychologica* 75, 1 (1990), 75–89.

David McNaughton, Janice Light, David R Beukelman, Chris Klein, Dana Nieder, and Godfrey Nazareth. 2019. Building capacity in AAC: A person-centred approach to supporting participation by people with complex communication needs. *Augmentative and Alternative Communication* 35, 1 (2019), 56–68.

W Scott Peavler. 1974. Pupil size, information overload, and performance differences. *Psychophysiology* 11, 5 (1974), 559–566.

John L Sibert, Mehmet Gokturk, and Robert A Lavine. 2000. The reading assistant: eye gaze triggered auditory prompting for reading remediation. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*. 101–107.

Olympia Simantiraki, Martin Cooke, and Simon King. 2018. Impact of Different Speech Types on Listening Effort.. In *INTERSPEECH*. 2267–2271.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.

Keiichi Tokuda, Heiga Zen, and Alan W Black. 2002. An HMM-based speech synthesis system applied to English. In *IEEE Speech Synthesis Workshop*. 227–230.

Cassia Valentini-Botinhao, Junichi Yamagishi, Simon King, and Yannis Stylianou. 2013. Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of HMM-based synthetic speech in noise.. In *Interspeech*. 3567–3571.

Matthew B Winn, Dorothea Wendt, Thomas Koelewijn, and Stefanie E Kuchinsky. 2018. Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in hearing* 22 (2018), 2331216518800869.

Alfred L Yarbus. 2013. *Eye movements and vision*. Springer.

Heiga Zen, Alan Senior, and Martin Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *Proc. ICASSP*. 7962–7966.