Modelling Attention Levels with Ocular Responses in a Speech-in-Noise Recall Task

Mateusz Dubiel
University of Luxembourg
Esch-sur-Alzette, Luxembourg
mateusz.dubiel@uni.lu

Minoru Nakayama[*]
Tokyo Institute of Technology
Tokyo, Japan
nakayama@ict.e.titech.ac.jp

Xin Wang
National Institute of Informatics
Tokyo, Japan
wangxin@nii.ac.jp

ABSTRACT

We applied state-space modelling technique to estimate the cognitive workload of a speech-in-noise (SIN) recall task, based on participants' oculo-motor responses to speech signals. We estimated common latent attention levels in 15 time bins and observed temporal changes between pupillary dilations and saccade frequencies, given that the both conditions were independent. We also compared two speech type factors (natural vs. synthetic) and three levels of signal-to-noise (-1dB, -3dB, and -5dB) using the estimated parameter distribution. The comparison of experimental factors provided us with insights into differences in participants' processing of spoken information during a SIN recall task.

CCS CONCEPTS

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

KEYWORDS

Statistical modelling, Cognitive load, Speech perception, Eye movement

ACM Reference Format:
Mateusz Dubiel, Minoru Nakayama, and Xin Wang. 2023. Modelling Attention Levels with Ocular Responses in a Speech-in-Noise Recall Task . In *2023 Symposium on Eye Tracking Research and Applications (ETRA '23), May 30–June 02, 2023, Tubingen, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3588015.3589665

1 INTRODUCTION

Since pupillary changes can be used as indices of task-evoked cognitive load [Beatty 1982; Zekveld et al. 2010], pupillometry has recently been applied to evaluation of text-to-speech (TTS) systems [Govender and King 2018; Govender et al. 2019; Winn et al. 2018]. We applied this technique to measure differences in cognitive workload between natural and synthetic speech (see [Dubiel et al. 2021b] for detailed analysis of the results). In this study, in addition to pupil dilations, saccade frequencies were also analysed as another index of mental workload. Since the characteristics of both indices are different, here, their relationships are analysed and discussed separately. The purpose of the current analysis is to extract the latent level of attention or mental workload during the experimental task. These levels have been estimated based on observed oculo-motor indices. In particular, the analysis focuses on exploring causal relationships of these responses, following the method outlined in [Nakayama and Hayakawa 2021].

Recently, several statistical modelling techniques were proposed to estimate users' latent activity level in cognitive tasks. One example of such technique is Bayesian modelling [Okano and Nakayama 2022; Ueno and Nakayama 2021] which can potentially be applied to estimate latent temporal changes of mental attention level from observed oculo-motor indices. The benefit of Bayesian modelling is that it allows for hierarchical decomposition of temporal data components and account for their individual contributions. For instance, in [Ueno and Nakayama 2021] hierarchical Bayesian modelling was applied to estimate stimulus detection response and microsaccade frequency in a dual detection task. This technique was also used to assess the relationship between temporal changes of individual ocular-motor metrics on cognitive workload during figure counting task [Okano and Nakayama 2022].

Here, we examine the feasibility of applying Bayesian modelling to a speech-in-noise (SIN) recall task, based on the experimental results discussed in [Dubiel et al. 2021a]. Specifically, we focus on the analysis of two types of speech (natural vs. synthetic) and three levels of signal-to-noise (-1dB, -3dB, and -5dB).

In sum, in this paper, we addresses the following two topics:

1. The feasibility of extracting temporal attention levels using pupillary changes and saccade frequencies during a SIN recall task by applying Bayesian modelling [Lee 2011] .
2. Correspondingly, based on the modelling results, we discuss the impact of experimental factors on estimated participants' attention.

Contrary to previous relevant work which explored modelling users' cognitive workload in visual perception and reasoning tasks [Okano and Nakayama 2022; Ueno and Nakayama 2021], here we focus on an auditory task. We decided to apply Bayesian modelling technique to a SIN recall task because it involves the use of working memory which requires attention, concentration and effort [Pichora-Fuller 2007; Rabbitt 1968; Rönnberg et al. 2013]. Specifically, as stipulated by The Ease of Language Understanding framework [Rönnberg et al. 2013], identifying speech masked by noise requires development and deliberate allocation of additional cognitive resources to encode the degraded speech, and match it with the listener's internal language representation.

Here, we seek to gain a better understanding of how different types of speech and levels of sound individually affect allocation of effort required for listening to and recalling speech samples.

[*] Corresponding author

ETRA '23, May 30–June 02, 2023, Tubingen, Germany

ACM ISBN 979-8-4007-0150-4/23/05.
https://doi.org/10.1145/3588015.3589665

2 EXPERIMENT SUMMARY

Below we provide a brief overview of the experimental procedure (please see [Dubiel et al. 2021a] for a more detailed explanation).

Speech stimuli consisted of forty travel-related speech samples that were selected from Japanese speech corpus of Saruwatari-lab., the University of Tokyo (JSUT) [1]. The selected samples were sampled at 48kHz and synthesised using a Japanese TTS system based on statistical parametric speech synthesis framework [Zen et al. 2013]. The selected samples were then mixed with a speech-shaped-noise at three levels of signal-to-noise (i.e., -1dB, -3dB, -5dB).

The experimental procedure is presented in Figure 1. Participants were asked to look at the black cross on a grey background, listen to speech samples (one at the time) and repeat the words that they heard when the cross changed its colour to red. Presentation of each stimuli consisted of five phases (calibration (1), listening (2), retention (3), recall (4) and relax (5)). Masking speech-shaped-noise was present until the recall phase. Participants were randomly assigned into one of three signal-to-noise levels and listened to blocks of stimuli (natural and synthetic) presented in a random order. The recall attempt was considered as 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 was processed as a repeated measure for each participant.

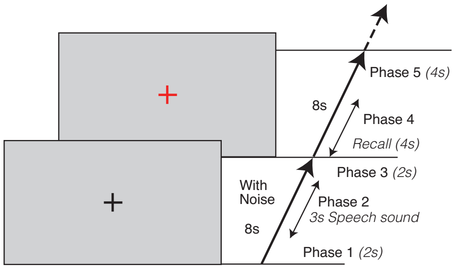


Figure 1: An overview of the experimental procedure.

The experiment involved sixteen native Japanese speakers (14 males and 2 females) with no self-reported hearing problems. The participants were aged between 21 and 25 years (mean = 22.5). All subjects attended a briefing session before the experiment and provided their consent in order to participate.

2.1 Recall accuracy

The recall accuracy was evaluated by a member of the research team who listened to participants' responses and classified them into 'correct' and 'incorrect' categories. The result indicated that the recall accuracy was significantly lower for synthetic speech than for natural speech at every signal-to-noise level except -3dB. In the current paper, the analysis provided in the subsection 2.2 and subsection 2.3 is based on the correct recall responses.

2.2 Pupil sizes

Temporal changes of pupil sizes for natural and synthetic speech stimuli are summarised in Figure 2. The pupil size is standardised individually based on the mean of pupil sizes from phase-1. As can be seen, pupil size increases until phase-3 and decreases from thereon until phase-5. The observed pupillary changes are in line with the results of the previous studies [Ahmed et al. 2023; Beatty 1982; Kahneman and Beatty 1966], where similar behaviour was observed during memorisation tasks.

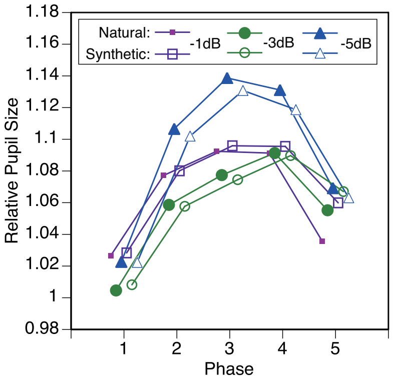


Figure 2: Pupillary changes for 5 experimental phases.

2.3 Saccade frequency

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 phase. The overall occurrence rate of saccades is summarised in Figure 3. Here, saccade occurrence rate corresponds to the proportion of a phase that contains saccades. Saccade occurrence decreases 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 the trial. Since the retention of information requires a high level of concentration [Andrewes 2015], saccade eye movements are suppressed in the middle of the trial while participants try to memorise what they heard. As can be observed, during phase-3, there are some differences between the conditions.

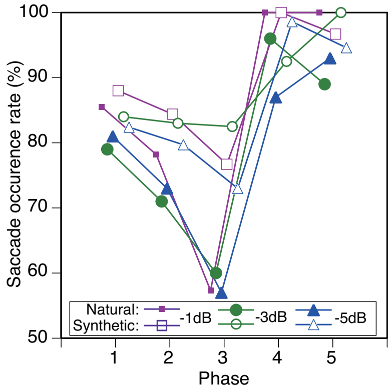


Figure 3: Rates of saccade occurrence

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

3 STATE-SPACE MODELLING PROCEDURE

The participants' cognitive behaviour is presented as a temporal attention level which contains a latent temporal sensing level using a Bayesian modelling technique [Haaf and Rounder 2017; Lee 2011], since the observation data is limited to a certain number of participants. In the following, we present two types of models of temporal changes for pupil size and saccade frequency (i.e., *state models* and *observation models*). We hypothesise that all factors in the models contribute independently to cognitive load of the participants. In particular, the attention levels are represent using two variables such as, $Attn$ as overall changes including experimental factors, and *S_level* as a temporal change during a task.

Here, the parameters are defined as, $Attn$: Temporal Attention levels over the observation as a source for ocular responses (pupil size/saccade rate) which is noted in a summation of following factors, *S_level*: latent temporal sensing level during the task as a temporal state model, *rID*: a factor of individuals, $Voice\_factor$: two-dimensional factor of Natural/Synthetic voices, *Sound_level*: three-dimensional factor of Sound levels (i.e., -1dB, -3dB, -5dB), *rPN*: 10-dimensional factor of presentation order (1~10), mu_{noise}: attention level ($Attn$) with an observation noise. Independent parameters are hypothesised to follow a normal distribution.

The model presentation consists of state models and observation models that generate both temporal pupil size and saccade frequency using the temporal attention level ($Attn$) driven by latent temporal sensing level (*S_level*), as shown in following equations.

- Common attention level

$$\begin{aligned} Attn = inv\_logit(S\_level + rID) \\ + Voice\_factor + Sound\_level + rPN \end{aligned} \tag{1}$$

State Model:

$$S\_level_i \sim normal(S\_level_{i-1}, \sigma_s)$$

- Pupil Diameter

Observation Model:

$$Pupil_{size} \sim normal(Attn, \sigma_p)$$

- Saccade frequency

$$\lambda = exp(mu_{noise})$$

Observation Model:

$$mu_{noise} \sim normal(Attn, \sigma_{noise})$$

$$NSac_{times} \sim poisson(\lambda)$$

In order to extract all the parameters in the above equations, the observed temporal data (i.e.,pupillary changes and saccade frequencies) was applied to a sampling technique of Markov Chain Monte Carlo (MCMC) procedure [Haaf and Rounder 2017; Lee 2011] with R and Stan packages. The converged models were obtained with 2000 steps with 500 burn-in lengths sampling in four chains using a fitness index such as $\hat{R}$ and Watanabe-Akaike Information Criterion (WAIC) [Watanabe and Opper 2010]. The number of latent temporal sensing levels during the task was controlled for both pupil and saccade indices. The following results are based on a converged solution is obtained at 15 temporal time bins as a result of sampling.

4 RESULT

4.1 Temporal change of attention sensing level

The estimated attention sensing parameters (*S_Level*) in 15 temporal changes are illustrated with parameter distribution with 1500 samplings in Figure 4 for pupil size (Fig 4 (a)) and for saccades (Fig 4 (b)). The small horizontal ticks indicate the mean of the distribution and the painted (maroon) regions correspond to the 95% confidence intervals (CI).

Initially, the attention sensing levels appear to be a common factor for both pupil dilation and saccades in Figure 4 ((a) and (b)). However, since later on these two conditions begin to diverge, another estimation is made by using the pupil response to vertically reflect wave-forms, transformed as (2 - $Pupil_{size}$) in Figure 4 (c).

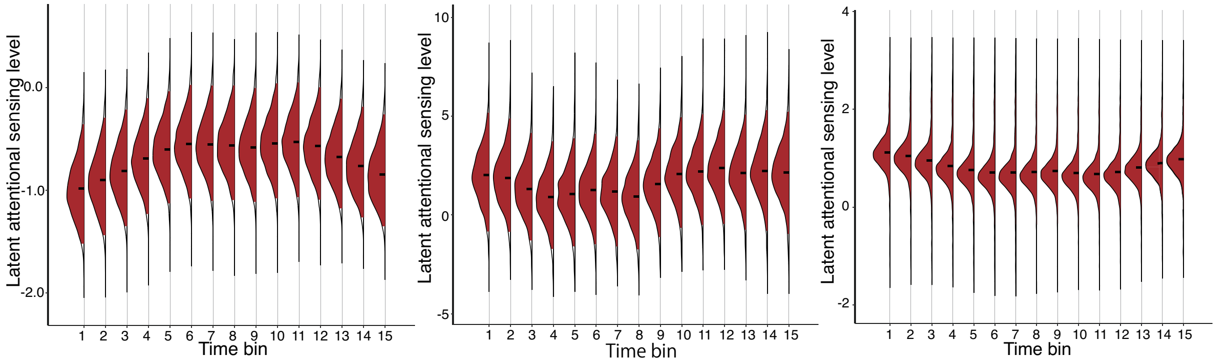
As can be observed, the local minima of attention sensing levels are almost synchronised. This modified response is used for the following analysis. Here, "*S_Level*" indicates the "remaining level of available attention resource". During the task, the available attention resources are reduced by the cognitive workload. In Figure 4 ((b) and (c)), the attention resource decreases due to increase of latent cognitive workload in the middle of session, and then increases at the end of the trial. Although the estimated level from both pupil and saccade responses shows a local peak in the middle, the overall trend of temporal changes differs between the two metrics. On the other hand, as illustrated in Figure 4 (c), the attention sensing level is minimised during the phase-2 of the observation. The cognitive workload is then released after the speech sample has been committed to a participant's memory.

4.2 Speech sound factor

The estimated parameters for speech sound factor $Voice\_factor$ are illustrated with parameter distribution in Figure 5 for pupils (Fig.5 (a)) size and saccades (Fig.5 (b)). Although there is little difference between means for pupil size, there is a substantial difference in saccades between natural and synthetic speech condition. We posit that the estimated distribution may correspond to the different impact of these two speech factors. The second, slightly smaller, peak in pupil size may indicate interaction between the factor of sound levels, which will be discussed in the following section.

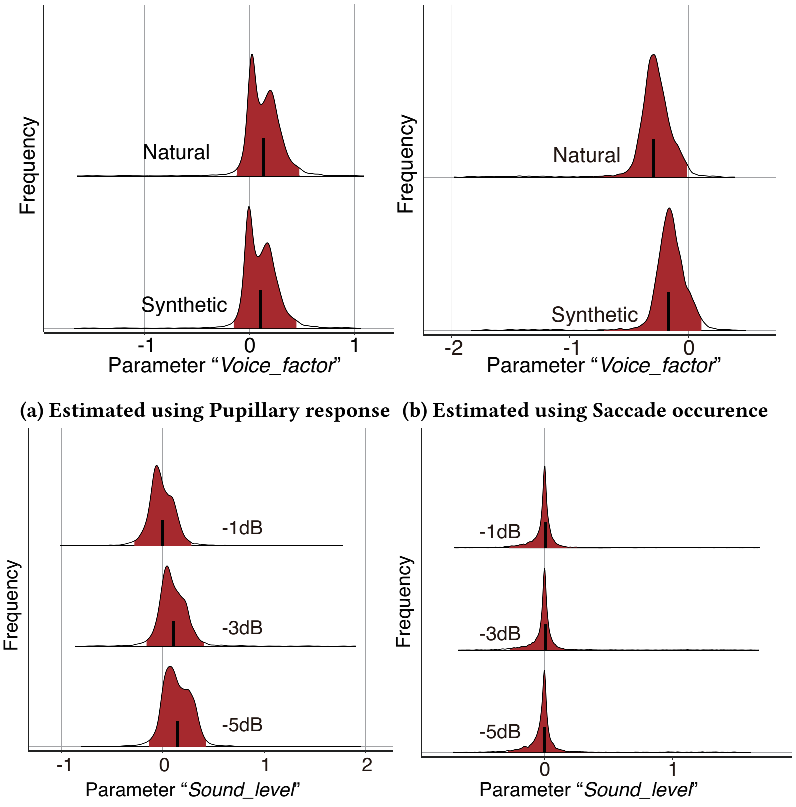
4.3 Sound level factor

The distributions of the parameter *Sound_level* are illustrated in Figure 5 for pupil size (Fig.5 (c)) and saccade frequency (Fig.5 (d)). For pupil size, means are gradually shifted to larger values as the signal-to-noise ratio decreases. Therefore, it could be postulated that the level of sound may contribute the attention sensing level. As for the $Voice\_factor$, secondary peaks can also be observed (albeit at a lower scale) — potentially indicating interaction between the two factors (i.e., speech type and sound level). For saccade rates, the mean level for -1dB is larger than the other two levels, while for the levels of -3dB and -5dB differences are marginal. Since the distributions for three levels are biased, the effect of individual sound levels will be examined closely in the following section.



(a) Pupillary **(b) Saccades** **(c) Pupillary reversed**

Figure 4: Distributions of the estimated "Latent attention sensing level (*S_level*)" using different ocular indices.



(a) Estimated using Pupillary response **(b) Estimated using Saccade occurence**

(c) Estimated using Pupillary response **(d) Estimated using Saccade occurence**

Figure 5: Summary of ocular responses for speech type factor (a,b) and noise level factor(c,d)

4.4 Hybrid model

As shown in Figure 4 (b) and (c), the estimated levels of latent attention sensing are slightly different and are both affected by the observed metrics. We applied another model to estimate the common level of attention sensing for both metrics with a higher precision. The iteration required 6000 steps and 500 burn-in step samplings to extract the same 15 bins. Distributions of the latent sensing levels in 5500 samples over the 15 bins are displayed using the same format as in Figure 6. The mean levels of the distributions show the temporal change, where the lowest bin is located around the middle of the trial (around time bin 6) which could be influenced by the task workload. This indicates that the estimated distributions reflect the overall workload instead of two types of estimated levels such as Figure 4 (a) and (b).

We also estimated other parameters in Equation 1 for the proposed hybrid model. The distributions for $Voice\_factor$ are shifted in two conditions (i.e., natural and synthetic) , and the distributions for $Sound\_level$ are shifted in three conditions (i.e., -1dB, -3dB, and -5dB). The sensing metric of oculo-motor indices for differentials of factors indicates the dominance on the parameter estimation in a hybrid model. The detailed analysis using the estimated attention levels is presented in section 5.

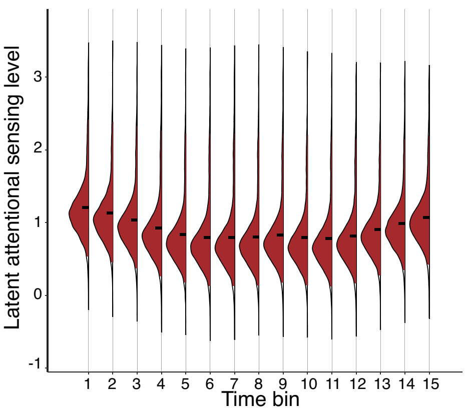


Figure 6: Distribution of the estimated "Attention sensing levels (S_level)" based on both pupil size and saccade frequency.

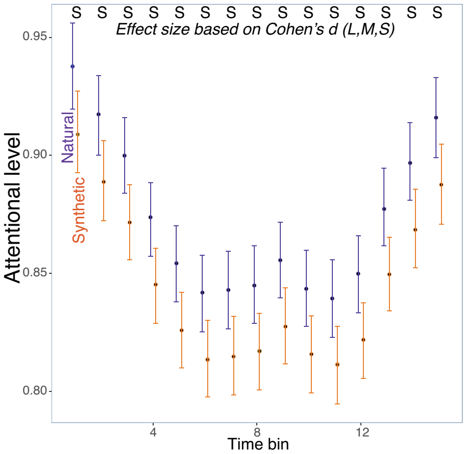


Figure 7: Estimated "Attention levels ($Attn$)" based on both pupil size and saccade frequency — summarised for both types of speech.

5 EVALUATION FOR ESTIMATED ATTENTION LEVELS

As shown in Equation 1, the attention level is calculated using the estimated parameters including individual factor (rID) and stimulus presentation order factor (rPN). Plausible attention levels ($Attn$) are generated for the sampled data using the trained model. The estimated attention level indicate "remaining attention resources" as well as the sensing level.

Means of the attention levels are calculated across the 15 bins for two speech conditions respectively. The results are summarised in Figure 7. The horizontal axis represents time bin, and the vertical axis represents the level of "remaining" attention. All plots are presented with the 95% confidence interval. As shown in Figure 7, there are two local minima which may indicate the end of the listening phase (around bin 6), and the beginning of recall (around bin 11). The difference between two speech types is tested using "Cohen's d", and the category of the effect size (S,M,L) [Anderson et al. 2017] is indicated for each bin in the top part of the graph. As can be seen in Figure 7 all differences between the two speech sound factors amount to a small effect size, with synthetic speech consistently leading to a higher reduction in the attention level. The exact values of Cohens'd and η^2 (Eta$\hat{2}$) are presented in Table 1.

Means of the attention levels for three sound levels are summarised in Figure 8 using the same format as in Figure 7. Means for the sound levels are illustrated from left to right for the -1dB, -3dB, and -5dB conditions. The "remaining" attention level is lowest for -1dB as the least noisy condition. Since the noise has been present during the first half of session, the difference is reduced in the second half. The attention levels for three SIN conditions are tested with One-way ANOVA, and the difference is tested using η^2 as effect size, and the contribution level (S,M,L) of test factor is extracted. Based on benchmarks suggested by Cohen [Cohen 2013], we use the following thresholds to refer to effect sizes; small (d = 0.2), medium (d = 0.5), and large (d ≥ 0.8). The effect size symbols are displayed in the top of graph such as "M-M" for two speech types.

The pairs of the first three bins and the last two bins in Figure 8 show a medium effect (M-M), and all the remaining pairs show a small effect (S-S). During the mid-stage of the task, the attention level is reduced by cognitive demands the task. Therefore, the ranges of three SIN levels are reduced, while the differences are large at the beginning and end of the session. The result shows the same levels of fluctuation between the two types of speech over the entire session.

6 DISCUSSION

The latent attention level during the SIN listening task is defined as a simple hierarchical model that involves several experimental factors, specified in Equation 1. In the current study, we examined the changes in latent attention level based on pupillary responses and saccade frequencies. We proposed two models for temporal changes (i.e., state model and observational model) and validated them for the observed oculo-motor responses. We also examined a hybrid model that combined the two constituent models mentioned above.

It should be noted that convergence of parameter estimation of predictive models depends on the number of sampling iterations and the number of time bins that represent temporal changes with time resolution. In particular, the larger the number of time bins and higher the complexity of model, the longer duration and larger number of iterations is required. We observed that the convergence time of parameter estimation gradually increased with the number

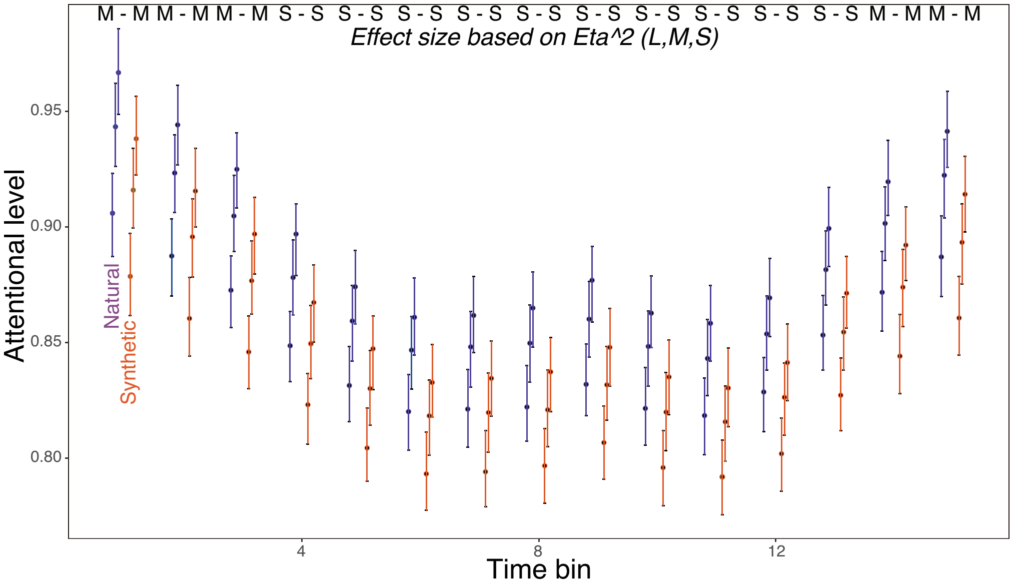


Figure 8: Estimated "Attention levels ($Attn$)" based on both pupil size and saccade frequency are summarised in two speech sounds and three noise levels. Note: means for the sound levels for each bin are illustrated from left to right for the -1dB, -3dB, and -5dB conditions.

Table 1: Comparison of effect sizes for synthetic and natural speech. Note: '*' indicates p < .05 , '' indicates p < .01, '***' indicates p < .001**

Comparison of Natural vs. Synthetic Speech

Time bin	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cohen's d	0.37	0.36	0.35	0.34	0.33	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.34	0.35
t (df = 312)	3.29^{**}	3.19^{**}	3.06^{**}	2.98^{**}	2.89^{**}	2.87^{**}	2.86^{**}	2.83^{**}	2.88^{**}	2.81^{**}	2.81^{**}	2.84^{**}	2.88^{**}	3.00^{**}	3.09^{**}

Natural sound level for one-way ANOVA

Time bin	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Eta^2	0.1	0.09	0.07	0.05	0.04	0.04	0.04	0.04	0.05	0.04	0.03	0.04	0.05	0.06	0.08
F (df=1,153)	17.84^{***}	14.41^{***}	11.18^{**}	8.74^{***}	6.49^{*}	5.82^{*}	5.74^{*}	6.47^{*}	7.32^{**}	5.89^{*}	5.47^{*}	5.80^{*}	7.95^{**}	8.91^{*}	12.48^{***}

Synthetic sound level for one-way ANOVA

Time bin	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Eta^2	0.1	0.08	0.07	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.04	0.05	0.06	0.08
F (df=1,158)	18.33^{***}	14.62^{***}	11.59^{***}	7.97^{**}	7.16^{**}	5.87^{*}	6.23^{*}	6.27^{*}	6.55^{*}	5.80^{*}	5.52^{*}	5.92^{*}	7.80^{**}	9.79^{**}	13^{***}

of time bins. Also the duration increased when the hybrid model was introduced: the iteration required 2000 MCMC steps for the composite model, and 6000 steps for the hybrid model to converge. The calculation duration for the hybrid model took three times longer than for each composite model. When the number of time bins was extended from 15 to 20 in (as illustrated in Figure 6), the duration increases 2.5 times. Therefore, representational ability of the model depends on the convergence availability and calculation duration. During the calculation for the hybrid model, pupil size was transformed to a flipped temporal change. When the original responses of pupil size were introduced, most of the estimated sensing levels were flattened indicating that two types of responses: pupil size and saccade frequency may be conflicted. To ensure high robustness the developed model needs to be able to accept various types of signal sources, thus a more flexible model description is required.

Estimated attention levels are summarised in Figure 7, there are significant differences in the levels between the two speech sounds (i.e, natural and synthetic). The differences between the speech types remain constant over the course of the session. As shown in Equation 1, speech category ($Voice\_factor$) is represent as a constant bias that may affect the overall attention levels. On the other hand, since contribution of noise level changes during the session as shown in Figure 8, the factor ($Sound\_level$) may have a simple bias towards the attention levels. In the Equation 1, all factors are represented as a linear summation. Although there is a possibility that described factors may interact with other unaccounted for factors, in the current analysis such plausible additional factors have been ignored. A more appropriate expression in a model equation should allow for flexible representation of reactions of oculo-motor indices in response to the experimental task workload. Such an expansion of the model will be a subject of our follow-up study.

Nonetheless, regardless of its apparent limitations, our proposed modelling technique can estimate latent attention levels during tasks that require focus and allocation of cognitive resources (such as SIN listening task). This technique can help to better understand the individual speech and non-speech (noise) factors on cognitive load induced by speech interfaces. For instance, since augmentative and assistive communication devices that feature speech have strong potential to increase social involvement of individuals with physical disabilities and complex information needs, it is important to better understand the cognitive implications that such devices have on users [McNaughton et al. 2019].

7 LIMITATIONS

We are mindful that our study is subject to several limitations. Firstly, given that our experimental data was labelled by only one person could have introduced an implicit bias. Secondly, since all of the participants were young adults with no self-reported hearing problems, our findings may not generalise to different age groups and users with different levels of hearing impairment. Therefore, in order to address these limitations, the future research should include more varied participant samples and involve multiple annotators to further improve the ecological validity.

8 SUMMARY

In this work, we estimated latent participants' attention levels using a statistical modelling technique with the MCMC sampling in order to extract internal information processing and mental workload during a speech-in-noise recall task. The observed oculo-motor indices used as cognitive load metrics were pupil sizes and saccade frequencies. Our proposed technique provides a way to model individual contributions of different sound factors (i.e., speech type and level of noise) and indicate their respective contributions towards cognitive workload in the SIN recall task.

REFERENCES

Arif Ahmed, Gondy Leroy, Han Yu Lu, David Kauchak, Jeff Stone, Philip Harber, Stephen A Rains, Prashant Mishra, and Bhumi Chitroda. 2023. Audio delivery of health information: An NLP study of information difficulty and bias in listeners. *Procedia Computer Science* 219 (2023), 1509–1517.

Samantha F. Anderson, Ken Kelly, and Scott E. Maxwell. 2017. Sample-Size Planning for More Accurate Statistical Power: A Met hod Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological Science* 28(11) (2017), 1547–1562.

David Andrewes. 2015. *Neuropsychology: From theory to practice.* Psychology Press.

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

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences.* Academic press.

Mateusz Dubiel, Minoru Nakayama, and Xin Wang. 2021a. Combining Oculo-motor Indices to Measure Cognitive Load of Synthetic Speech in Noisy Listening Conditions.. In *ACM Symposium on Eye Tracking Research and Applications.* 1–6.

Mateusz Dubiel, Minoru Nakayama, and Xin Wang. 2021b. 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.

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.

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

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

Julia M. Haaf and Jeffrey N. Rounder. 2017. Developing constraint in Bayesian mixed models. *Psychological Methods* 22 (2017), 779–798.

Daniel Kahneman and Jackson Beatty. 1966. Pupil diameter and load on memory. *Science* 154, 3756 (1966), 1583–1585.

Michael D. Lee. 2011. How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology* 55 (2011), 1–7.

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.

Minoru Nakayama and Yoshiya Hayakawa. 2021. Influence of Task-evoked Mental Workloads on Oculo-motor indices and their connections. *EAI Trans. Context-aware Systems and Application* 7, 23 (2021), e2:1–10.

Tomomi Okano and Minoru Nakayama. 2022. Research on Time Series Evaluation of Cognitive Load Factors using Features of Eye Movement. In *Proc. ETRA2022, COGAIN Workshop.* ACM, NY, USA, 61:1–6.

MK Pichora-Fuller. 2007. Audition and cognition: What audiologists need to know about listening. *Hearing care for adults* (2007), 71–85.

Patrick MA Rabbitt. 1968. Channel-capacity, intelligibility and immediate memory. *The Quarterly journal of experimental psychology* 20, 3 (1968), 241–248.

Jerker Rönnberg, Thomas Lunner, Adriana Zekveld, Patrik Sörqvist, Henrik Danielsson, Björn Lyxell, Örjan Dahlström, Carine Signoret, Stefan Stenfelt, M Kathleen Pichora-Fuller, et al. 2013. The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in systems neuroscience* 7 (2013), 31.

Takahiro Ueno and Minoru Nakayama. 2021. Estimation of Visual Attention using Microsaccades in response to Vibrations in the Peripheral Field of Vision. In *Proc. ETRA2021.* ACM, NY, USA, 19:1–6.

Sumio Watanabe and Manfred Opper. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research* 11, 12 (2010).

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.

Adriana A Zekveld, Sophia E Kramer, and Joost M Festen. 2010. Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and hearing* 31, 4 (2010), 480–490.

Heiga Zen, Andrew Senior, and Mike Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing.* IEEE, 7962–7966.