Interspeech 2018
2-6 September 2018, Hyderabad

Measuring the cognitive load of synthetic speech using a dual task paradigm

Avashna Govender, Simon King

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
a.govender@sms.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

We present a methodology for measuring the cognitive load (listening effort) of synthetic speech using a dual task paradigm. Cognitive load is calculated from changes in a listener's performance on a secondary task (e.g., reaction time to decide if a visually-displayed digit is odd or even). Previous related studies have only found significant differences between the best and worst quality systems but failed to separate the systems that lie in between. A paradigm that is sensitive enough to detect differences between state-of-the-art, high quality speech synthesizers would be very useful for advancing the state of the art. In our work, four speech synthesis systems from a previous Blizzard Challenge, and the corresponding natural speech, were compared. Our results show that reaction times slow down as speech quality reduces, as we expected: lower quality speech imposes a greater cognitive load, taking resources away from the secondary task. However, natural speech did not have the fastest reaction times. This intriguing result might indicate that, as speech synthesizers attain near-perfect intelligibility, this paradigm is measuring something like the listener's level of sustained attention and not listening effort.

Index Terms: cognitive load, dual task paradigm, speech synthesis

1. Introduction

Cognitive load is referred to as *listening effort* in the field of speech understanding and is defined as the amount of mental effort exerted to perform a listening task [1]. In other words, it is the amount of cognitive resources a listener allocates to the task of processing speech.

To quantify listening effort, the dual-task paradigm is commonly used. This is based on an assumption that human cognitive capacity is limited [2]. When the brain is forced to undertake a secondary task at the same time as a demanding primary task, the resources available for the secondary task are limited. As a consequence, performance on the secondary task deteriorates (assuming the primary task is prioritized). The amount of deterioration in performance on the secondary task is a behavioural measure that reflects the cognitive load of the primary task. Several studies have used this paradigm to investigate the effects of various conditions on the listening effort of natural speech, including noise [3, 4, 5, 6], listener age [6, 7, 8], and noise reduction algorithms [9].

There is much less research into the listening effort of synthetic speech. Many studies have used the reaction time paradigm to explore differences in perception of natural and synthetic speech, and only a few discuss the relationship to cognitive load [10, 11, 12, 13]. These studies consistently show that listeners respond faster when listening to natural speech than to synthetic speech. This suggests a delay in the processing of synthetic speech. Researchers in the field believe that this delay occurs in the early stages of speech perception that involve the recognition of words. It has been claimed that this delay is caused by an increased mental workload required to process synthetic speech. However, this research was done in the 1980s and 1990s using rule-based formant synthesizers. Subsequent developments in speech synthesis using waveform concatenation [14], then Hidden Markov Models (HMM) [15], and most recently Deep Neural Networks (DNN) [16, 17], mean that speech synthesizers now attain near perfect intelligibility. If increased cognitive load is caused simply by reduced intelligibility, then as synthesizers gain better intelligibility scores the listening effort should decrease.

In [13], comprehensibility of various speech synthesizers and a human voice were compared using the dual task paradigm. Significant differences were found only between the human voice and the worst quality system but not between the systems that lie in between. The study in [13] used the dual-task paradigm to measure intelligibility. But it is not actually clear exactly what this paradigm is measuring - intelligibility or listening effort. A drawback of the study was that the synthesizers compared were built using different speech data.

Since 2005, evaluating speech synthesizers with controlled speech data was made possible through the Blizzard Challenge [18]. By eliminating the variability of the data, much fairer comparisons across speech synthesizers can now be obtained. Therefore more meaningful results can now be obtained from the dual task paradigm than in [13].

Another issue with the dual-task paradigm is the wide variability in experimental designs for measuring listening effort, especially in the choice of secondary task [19]. Some studies find no differences amongst varying secondary tasks, whilst others do. [20] investigated the effect of changing the secondary task, comparing: a simple visual probe, a complex visual probe and a word-category recognition task. Results showed that only the word-category recognition task affected performance on the primary task. The investigators believe that the word-category recognition task required deeper processing than the two visual-probe tasks. However, they are uncertain whether it was more sensitive because both primary and secondary tasks were linguistic or because the word-category recognition secondary task was simply more demanding than the visual probes. The review in [19] concludes that further research is still necessary to identify what type of secondary task is best suited for investigating listening effort.

In light of the problems described above, there is an obvious need for a better methodology. With increased use of speech synthesizers in real world applications, especially those involving multi-tasking such as driving a vehicle, measuring the listening effort of synthetic speech is becoming more important. By developing a paradigm that is sensitive enough to detect differences between state-of-the-art, high quality speech synthesizers, we hope to gain a better understanding on how synthetic speech interacts with the human cognitive processing system. This will support the development of new methods for

10.21437/Interspeech.2018-1199

evaluation, therefore advancing the state-of-the-art. This paper presents a dual-task methodology which attempts to address some of the above problems. In our study, speech synthesizers from a previous Blizzard Challenge and the corresponding natural speech were compared, in two experiments using differing secondary tasks.

2. Experimental design

2.1. Tasks

The primary task chosen for this study was sentence recognition. The listener was instructed to listen to a spoken audio sample and verbally repeat the sentence as accurately as possible. The secondary task for Experiment 1 was a visual motor task where the listener needed to decide if a visually-displayed digit is odd or even and press the appropriate button on a response box. The secondary task for Experiment 2 was inspired by [20] who suggest that a word-category secondary task is sensitive to listening effort. The odd-even digit task of Experiment 1 was therefore replaced with a lexical decision task for Experiment 2. Listeners were asked to decide whether a visually-displayed word exists or not. All words had the same number of characters. The motivation is that this secondary task requires the listener to make use of the same (limited) cognitive resources as the primary task, which should result in a greater cognitive load.

2.2. Structure

Experiments 1 and 2 had the same three sections: (1) secondary task alone; (2) dual-task practice; (3) dual-task.

Section 1 comprised 11 blocks. The first of the 11 blocks was a practice block of 5 trials and the remaining 10 blocks were constructed using a 5 x 5 Latin square repeated twice to balance listeners and systems (and, in section 3, sentences). In each trial, two digits (or words, in Experiment 2) were displayed sequentially. The first was displayed a random time delay after the onset of the trial. The second immediately followed the listener's decision response to the first. Listeners were instructed to respond as quickly as possible.

To mitigate the training effect from section 1 (the baseline) to section 3 (dual-task), listeners were only allowed to proceed after 85% accuracy was achieved in section 1.

Section 2 just had one block of 3 trials, all using natural speech, to familiarise the listener with the dual-task whilst avoiding exposure to the synthetic speech to be heard in section 3. The listener could repeat section 2 as many times as he or she wished.

Section 3 had identical structure to section 1, the only difference being that listeners also had to listen to and repeat a sentence. The first (practice) block used one audio sample from each of the systems in Table 1. Each of the subsequent blocks used five audio samples all from only one of the five systems, with each system appearing in two blocks. All audio samples were 2.5–3s in duration. In section 3, listeners were instructed to prioritize the listening-and-repeating (primary) task over the decision (secondary) task. Their verbal responses were recorded in order to confirm that they were correctly prioritising the primary task. Self-report measures were also taken: the listener was asked to rate the naturalness of each audio sample and the difficulty of listening to it, each on 5-point scales labelled *1 - unnatural* to *5 - natural*, and *1 - very difficult* to *5 - very easy* respectively.

2.3. Participants and listening conditions

For each experiment, 20 paid native English speakers were recruited, ranging in age from 18 to 28, with no self-reported hearing problems.

Each sentence was played diotically to the listeners using Beyerdynamic DT770 headphones in individual soundproof booths. Stimuli were presented using E-Prime 2.0 software [21]. Reaction time on the decision (secondary) task and the self-report responses were recorded with an E-Prime response box and their verbal responses were recorded using a microphone.

2.4. Speech materials

Stimuli presented to the participants were sentences generated by four synthesizers taken from the 2011[1] Blizzard Challenge [22] and natural versions from the same speaker. All systems were created using approximately 16.6 hours of speech from a US English female professional speaker.

The dual-task paradigm requires that all available cognitive resources are used: differences in listener performance on the secondary task will only be observed if the total cognitive load on the listener exceeds capacity. We used clean speech, and so were concerned that this imposes insufficient load on the listener: listening to natural speech in ideal conditions is effortless [23]. In [19], degraded quality and sentence complexity are both reported to increase listening effort for natural speech. Therefore we opted to use Semantically Unpredictable Sentences (SUS) in the primary task to increase load. A summary of the selected speech synthesizers is in Table 1.

Table 1: *Summary of selected speech synthesis systems, with their scores from the Blizzard Challenge 2011 for naturalness (Mean Opinion Score, MOS – higher is better) and intelligibility (Word Error Rate, WER – lower is better)*

System	MOS median (mean)	WER %
Natural	5 (4.8)	16
Hybrid	3 (3.3)	20
Unit Selection	3 (3.1)	22
HMM	3 (2.6)	20
Low-Quality HMM	1 (1.4)	26

2.5. Analysis

Listening effort is calculated as the difference in a listener's performance (where performance is defined as Reaction Time, RT) on the secondary task between the baseline and dual-task condition. If the listener optimizes performance in the primary task, it is assumed that the listener will perform equally well for both conditions in that primary task.

Only RTs where participants made the correct decision were considered during the analysis. Outliers of RTs less than 100ms or longer than 2 seconds were discarded. In sections 1 and 3, the first of the 11 blocks was a practice and responses were discarded, with results being calculated from the remaining 10 blocks. Listening effort was calculated using the propor-

[1] because our experimental design uses naturally-spoken Semantically Unpredictable Sentences, which are not available in later years

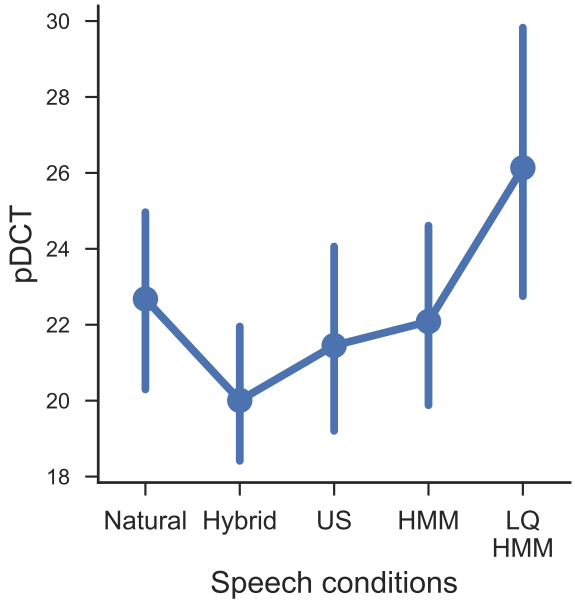

Figure 1: *Mean proportional dual cost time (pDCT) for Experiment 1*

tional dual cost time (pDCT) [19]:

$$pDCT = \frac{Sd - Sb}{Sb} \times 100$$

where Sb is the RT for the baseline condition, and Sd is the RT for the dual-task condition.

To test for significance a mixed analysis of variance (ANOVA) with repeated measures was applied. An alpha level of 0.05 was used for all statistical tests. If significance was observed, the paired t test with Bonferroni correction was used to determine which pairs of conditions had statistically significant differences.

3. Results

3.1. Experiment 1: Digit Task

Secondary task accuracy did not vary much across listeners and the system quality had no significant effect. Only 6.2% of responses were incorrect. The mean proportional dual cost time is presented in Figure 1 and – across synthesizers – is lowest for the Hybrid system and highest for the Low-Quality HMM voice, as expected, although no differences are statistically significant (ANOVA for repeated measures, F(4,76) = 1.75 with p = 0.14 ($p \leq 0.05$), n^2=0.05). This result suggests that, for highly intelligible speech synthesizers, differences in listening effort are difficult to detect.

The mean pDCT for natural speech is unexpected. If intelligibility was the main contributing factor, this should be lowest of all. Even though differences are not significant, this is nevertheless intriguing because it *suggests* that listeners respond faster (on the secondary task) when listening to the highest quality speech synthesizer than when listening to natural speech. The mean pDCT, at least across synthesizers, actually correlates (negatively) more with naturalness as opposed to intelligibility (both were measured in the Blizzard Challenge).

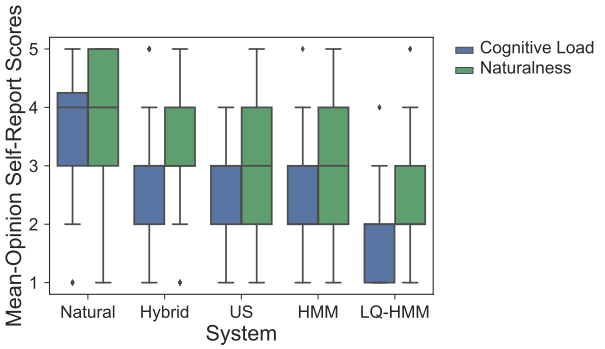

Figure 2: *Self-reported measures for Experiment 1*

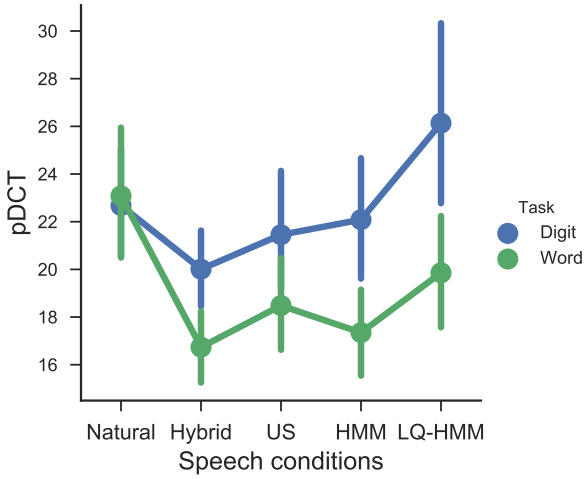

Figure 3: *Mean proportional dual cost time (pDCT) for Experiments 1 and 2*

The self-reported measures are presented in Figure 2 and show that the subjectively easiest system to listen to was natural speech and the most difficult was the Low-Quality HMM. Listeners reported no differences in difficulty amongst the higher quality synthesizers. It is surprising that listeners thought natural speech was the easiest to listen to, yet responded slower on the secondary task than when listening to the high quality Hybrid synthesizer. Listeners unsurprisingly found the human voice most natural followed by the hybrid voice, in line with findings from the Blizzard Challenge itself.

3.2. Experiment 2: Word Task

This experiment was devised in response to the intriguing result for natural speech in Experiment 1, replacing the digit task with a linguistic task. Once again, secondary task accuracy did not vary much across listeners and the system quality had no effect on accuracy. Only 8.05% of responses were incorrect, quite similar to performance on the digit task in Experiment 1.

Mean pDCT results are presented in Figure 3 and show that listeners responded faster on the word task than the digit task when listening to synthetic speech, whereas for natural speech their mean pDCT remained much the same. Using a mixed model ANOVA for repeated measures, F(4,76) = 5.01 with p = 0.001 ($p \leq 0.05$), n^2 = 0.09, we found significant differences

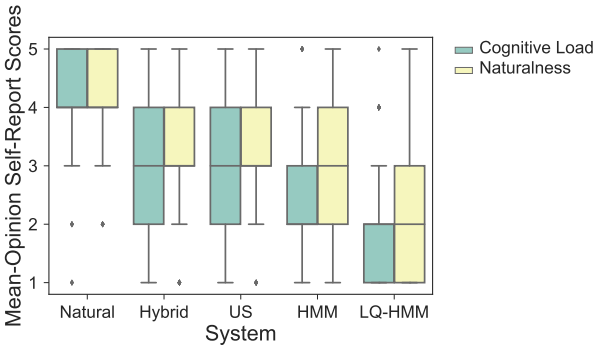

Figure 4: *Self-reported measures for Experiment 2*

in the results. Using a post-hoc test, significance difference was only found between the natural speech and the Hybrid synthesizer which are the systems with fastest and slowest mean pDCT. Once again, pDCT is on average fastest for the Hybrid system.

The trend in pDCT observed in Figure 3 now correlates more with the WER scores of the Blizzard Challenge, instead of naturalness as observed in the digit task. By changing from the digit task to the word task, it appears that the dual task paradigm shifts sensitivity from naturalness to intelligibility, for synthetic speech. Experiment 2 shows that the choice of secondary task influences how cognitive load is divided between primary task and secondary tasks, but only in in the case of synthetic speech. Natural speech behaves unexpectedly in both cases, and furthermore the mean pDCT remains the same: it is not affected by the choice of secondary task.

The self-reported measures for this experiment are presented in Figure 4. As in the digit task, listeners found natural speech to be easiest to listen to and Low-Quality HMM the most difficult.

The naturalness scores again show the natural voice as the most natural. The Hybrid, Unit Selection and HMM voices were rated similarly and the Low-Quality HMM system was rated unnatural. Self-reported listening effort correlates with self-reported naturalness, as well as with naturalness as measured in the Blizzard Challenge.

4. Findings and discussion

Mean pDCT is lower for good quality synthetic speech than for natural speech Previous reaction time studies have consistently shown that listeners respond faster when listening to natural speech than synthetic speech. Yet in both experiments presented in this paper, results show the opposite effect. If this paradigm really is measuring listening effort, then our result suggests that listening effort for synthetic speech (in all cases except the lowest quality) might be lower than for natural speech. This seems unlikely, so perhaps a more plausible conclusion is that the paradigm is not actually measuring listening effort but something else, such as the listener's sustained level of attention.

For natural speech, pDCT was the same for both experiments Processing natural speech in quiet conditions is unaffected by the type of secondary task. The dual-task condition does increase RTs (i.e., pDCT is greater than zero) but this is presumed to indicate that the dual-task paradigm is not only measuring listening effort, at least in the case of natural speech.

For synthetic speech, pDCTs are unexpectedly lower for the linguistic secondary task (Experiment 2) than for the non-linguistic task (Experiment 1) Given that the primary task is linguistic, we expected the linguistic secondary task to be more affected than the non-linguistic task, but the reverse was found. Although the differences are not statistically significant, this unexpected result suggests that there is some other factor in play, which we do not yet understand.

For synthetic speech, in Experiment 1, as naturalness decreases, reaction time increases We speculate that the digit task is too easy, and therefore listeners' processing capacity is not fully utilised. This leaves spare resources that can be allocated to non-essential aspects of the listening task. As a consequence, the listener is able to process more of the detailed acoustic cues of synthetic speech and make a judgement about naturalness. On the other hand, for lower quality synthetic speech more resources could be utilised due to inconsistent or missing acoustic cues.

For synthetic speech, in Experiment 2, as intelligibility decreases, reaction time increases In contrast to the digit task, the linguistic (word) secondary task, in our opinion, fully utilises all available resources, given that the listener must use the same cognitive resources for both the primary and secondary task. Therefore, no spare resources remain available for processing non-essential acoustic details. With only limited resources available, the listener can only execute the essential task of decoding the words.

5. Conclusion

In conclusion, measuring listening effort of synthetic speech remains a challenge. In situations where the cognitive load imposed by a secondary task is sufficiently high, speech synthesizers with lower intelligibility require higher listening effort. This has implications for real-world multi-tasking applications such as speech synthesis heard by vehicle drivers.

On the other hand, when speech reaches ceiling intelligibility, like natural speech for example, the dual task paradigm appears to be measuring something else, possibly the level of sustained attention. As speech synthesizers are indeed approaching near-perfect intelligibility (albeit in quiet conditions only, thus far), this points to new methods for evaluation.

Finally, it is clear that the choice of secondary task in the dual-task paradigm is very important. Different aspects of the synthetic speech appear to have an effect, depending on the type of task chosen, notably linguistic vs. non-linguistic. In ongoing work we are investigating pupillometry as a more sensitive measure of listening effort [24].

6. Acknowledgements

This project has received funding from the EUs H2020 research and innovation programme under the MSCA GA 67532 (the ENRICH network: `www.enrich-etn.eu`).

7. References

[1] R. McGarrigle, K. J. Munro, P. Dawes, A. J. Stewart, D. R. Moore, J. G. Barry, and S. Amitay, "Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group white paper," *International journal of audiology*, vol. 53, pp. 443–440, 2014.

[2] D. Kahneman, *Attention and effort*. Prentice-Hall Englewood Cliffs, NJ, 1973, vol. 1063.

[3] S. Fraser, J.-P. Gagné, M. Alepins, and P. Dubois, "Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 1, pp. 18–33, 2010.

[4] C. S. Howard, K. J. Munro, and C. J. Plack, "Listening effort at signal-to-noise ratios that are typical of the school classroom," *International journal of audiology*, vol. 49, no. 12, pp. 928–932, 2010.

[5] B. W. Hornsby, "The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands," *Ear and Hearing*, vol. 34, no. 5, pp. 523–534, 2013.

[6] J. L. Desjardins and K. A. Doherty, "Age-related changes in listening effort for various types of masker noises," *Ear and hearing*, vol. 34, no. 3, pp. 261–272, 2013.

[7] P. A. Gosselin and J.-P. Gagné, "Older adults expend more listening effort than young adults recognizing audiovisual speech in noise," *International journal of audiology*, vol. 50, no. 11, pp. 786–792, 2011.

[8] P. A. Tun, S. McCoy, and A. Wingfield, "Aging, hearing acuity, and the attentional costs of effortful listening." *Psychology and aging*, vol. 24, no. 3, p. 761, 2009.

[9] A. Sarampalis, S. Kalluri, B. Edwards, and E. Hafter, "Objective measures of listening effort: Effects of background noise and noise reduction," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 5, pp. 1230–1240, 2009.

[10] D. Pisoni, "Speeded classification of natural and synthetic speech in a lexical decision task," *The Journal of the Acoustical Society of America*, vol. 70, no. S1, p. S98, 1981.

[11] S. A. Duffy and D. B. Pisoni, "Comprehension of synthetic speech produced by rule: A review and theoretical interpretation," *Language and Speech*, vol. 35, no. 4, pp. 351–389, 1992.

[12] C. Delogu, S. Conte, and C. Sementina, "Cognitive factors in the evaluation of synthetic speech," *Speech Communication*, vol. 24, no. 2, pp. 153–168, 1998.

[13] G. P. Sonntag, T. Portele, and F. Haas, "Comparing the comprehensibility of different synthetic voices in a dual task experiment," in *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, Blue Mountains, Australia, 1998, pp. 5–10.

[14] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing (ICASSP), 1996 IEEE International Conference*, vol. 1. Atlanta, USA: IEEE, 1996, pp. 373–376.

[15] H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A. W. Black, and K. Tokuda, "Recent development of the HMM-based speech synthesis system (HTS)," in *APSIPA*, Sapporo, Japan, 2009, pp. 121–130.

[16] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference*. IEEE, 2013, pp. 7962–7966.

[17] T. Merritt, R. A. Clark, Z. Wu, J. Yamagishi, and S. King, "Deep neural network-guided unit selection synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference*. Lujiazui, Shangai: IEEE, 2016, pp. 5145–5149.

[18] K. Tokuda and A. W. Black, "The Blizzard Challenge-2005: Evaluating corpus-based speech synthesis on common datasets," in *Interspeech*, Lisbon, Portugal, 2005, pp. 77–80.

[19] J.-P. Gagne, J. Besser, and U. Lemke, "Behavioral assessment of listening effort using a dual-task paradigm: A review," *Trends in hearing*, vol. 21, pp. 1–25, 2017.

[20] E. M. Picou and T. A. Ricketts, "The effect of changing the secondary task in dual-task paradigms for measuring listening effort," *Ear and Hearing*, vol. 35, no. 6, pp. 611–622, 2014.

[21] W. Schneider, A. Eschman, and A. Zuccolotto, *E-Prime User's Guide*, Psychology Software Tools, Inc., Pittsburgh, 2012.

[22] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Blizzard Challenge*, Florence,Italy, 2011.

[23] J. Rönnberg, T. Lunner, A. Zekveld, P. Sörqvist, H. Danielsson, B. Lyxell, Ö. Dahlström, C. Signoret, S. Stenfelt, M. K. Pichora-Fuller *et al.*, "The ease of language understanding (ELU) model: theoretical, empirical, and clinical advances," *Frontiers in systems neuroscience*, vol. 7, p. 31, 2013.

[24] A. Govender and S. King, "Using pupillometry to measure the cognitive load of synthetic speech," in *Interspeech*, Hyderabad, India, 2018.