INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Using pupil dilation to measure cognitive load when listening to text-to-speech in quiet and in noise

Avashna Govender[1], *Anita E. Wagner*[2], *Simon King*[1]

[1]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2]Department of Otorhinolaryngology / Head and Neck Surgery, University Medical Center Groningen, University of Groningen, The Netherlands

a.govender@sms.ed.ac.uk, a.wagner@umcg.nl, Simon.King@ed.ac.uk

Abstract

With increased use of text-to-speech (TTS) systems in real-world applications, evaluating how such systems influence the human cognitive processing system becomes important. Particularly in situations where cognitive load is high, there may be negative implications such as fatigue. For example, noisy situations generally require the listener to exert increased mental effort. A better understanding of this could eventually suggest new ways of generating synthetic speech that demands low cognitive load. In our previous study, pupil dilation was used as an index of cognitive effort. Pupil dilation was shown to be sensitive to the quality of synthetic speech, but there were some uncertainties regarding exactly what was being measured. The current study resolves some of those uncertainties. Additionally, we investigate how the pupil dilates when listening to synthetic speech in the presence of speech-shaped noise. Our results show that, in quiet listening conditions, pupil dilation does not reflect listening effort but rather attention and engagement. In noisy conditions, increased pupil dilation indicates that listening effort increases as signal-to-noise ratio decreases, under all conditions tested.

Index Terms: text-to-speech, cognitive load, pupillometry, adverse conditions

1. Introduction

Evaluation methods have remained much the same for text-to-speech synthesis despite increasingly-diverse real-world applications. This is concerning: there may be potential negative implications when listening to synthetic speech, especially in adverse conditions. Specifically, evaluation methods generally fail to consider the *cognitive load* imposed by listening to synthetic speech.

[1] compared listening effort across different speech types in the presence of speech-shaped noise, showing that synthetic speech demanded the greatest effort even at favourable signal-to-noise ratios. This highlights the impact synthetic speech has on the human cognitive processing system. It is necessary to better understand these impacts and eventually to develop synthetic speech that imposes a lower load on listeners.

In our previous study [2], we used pupillometry to measure the cognitive load of synthetic speech. The choice of pupillometry was motivated by several studies that consistently showed a correlation between pupil dilation and the mental effort required to carry out a specific task [3, 4, 5, 6, 7]. The results in [2] showed that pupil dilation was sensitive to the quality of synthetic speech. Differences were observed between natural speech and synthetic speech whilst differences between several speech synthesizers were much more difficult to detect.

It was not possible to tell whether cognitive load (CL) truly was equivalent across all TTS systems compared, or whether the task was not cognitively demanding enough (due to listening in quiet). Furthermore, ANOVA with repeated measures was used as the primary statistical test. According to Mirman [8], tests such as ANOVA are not well-suited for time-series data, whilst tests such as growth curve analysis (GCA) offer advantages for time-series data. With GCA the entire time-course of the data is analysed as opposed to binned data used in ANOVA that results in a lot of meaningful information being lost.

To address these concerns, we re-analyse experiment 3 from [2] using growth curve analysis. This is experiment 1 in this paper. The results are then compared with two new experiments involving listening to synthetic speech in the presence of speech-shaped noise: experiments 2 and 3 in this paper.

2. Experimental Design

2.1. Participants and Speech Material

30 native English speakers with no self-reported hearing problems, age 19 to 37 years, were recruited and divided evenly between experiments 2 and 3.

We used sentences generated by four synthesizers taken from the 2011 Blizzard Challenge [9] and natural versions from the same speaker. The synthesizers are: Hybrid, Unit Selection (US), Hidden Markov Model (HMM) and Low-Quality HMM (LQ-HMM). All were created from the same 16.6 hours of speech from a English female professional speaker with US accent. Since we wish to measure cognitive load of synthetic speech for real applications, meaningful sentences are used in all experiments, taken from the Glasgow Herald newspaper.

As in [2], stimuli were blocked by system, resulting in 5 blocks, each containing 20 sentences. The block order was balanced using a 5x5 Latin square design to ensure all listeners, systems and sentences were equally represented. At the end of each block, self-reported cognitive load, motivation to listen, and naturalness scores were collected on 5-point rating scales.

2.2. Experimental set-up

The set-up of this experiment is the same as [2]. To summarize: the speech stimuli described earlier were played to listeners through headphones in a noise- and light-controlled room. Simultaneously, pupil size was monitored using an eye tracker. All stimuli were mixed with speech-shaped noise at signal-to-noise ratios (SNRs) -1dB and -3dB, chosen such that the cognitive effort is increased whilst intelligibility remains close to ceiling. In accordance with the estimated psychometric function in [10] which related keyword scores to SNR for speech-shaped

http://dx.doi.org/10.21437/Interspeech.2019-1783

Table 1: *Summary of interpretation of each time term in GCA Formula: ERPD ~(time1 + time2 + time3) * CONDITION + (time1 + time2 + time3 |SUBJECT) + (time1 + time2 + time3 |ITEM + (time1 + time2 + time3 |GROUP)*

Term	Interpretation
Intercept	Overall mean pupil dilation
Linear (time1)	Overall rate of pupil dilation
Quadratic (time2)	Shape of peak
Cubic (time3)	Falling slope

noise, the expected keyword correct percentages at -1dB and -3dB are approximately 80% and 60% respectively. For comparison, in [1], the TTS condition at -5dB SNR was too difficult.

The procedure described in [2] was followed exactly in terms of structure, presentation and data collection.

2.3. Pre-processing and analysis

After data collection, improvements to pre-processing and analysis were made by following the guidelines in [11]. The mean and standard deviations (SD) of the pupil size, from 1 second before sentence onset (baseline) until the start of the verbal response, were calculated. Pupil size values more or less than 2 SD to the mean were coded as blinks or artifacts. If total blink duration was more than 20% of the trial, or an individual blink was longer than 300ms, that trial was excluded. For retained trials, blinks were removed using linear interpolation using a window from 50 samples before the detected blink until 80 samples after. After deblinking, the data were downsampled to 50Hz for faster processing. Subsequently, the Event Related Pupil Dilation (ERPD) was computed. This was calculated using the equation in [12].

GCA was used to analyze the time course of the ERPD within a specific time period in which the peak was observed. The overall time course of the data was captured using a second-order (quadratic) or third-order (cubic) orthogonal polynomial with fixed effects of condition (various synthesizers) on all time terms. The participant, group (with respect to the Latin square design) and item (sentence stimulus) were used as random effects on all time terms. Post-hoc tests were performed by changing the baseline condition and cycling through each of the five conditions to get comparisons across all conditions for each time term. Table 1 summarizes what each time term represents. Statistical significance (p-values) for individual parameter estimates were assessed using the normal approximation (i.e., treating the t-value as a z-value). All analyses were carried out in R.

3. Results

3.1. Experiment 1: Quiet condition

The results for the quiet condition are obtained by re-analysis of the raw pupil data collected for Experiment 3 in [2] and are presented in Figure 1, which looks slightly different from Figure 6 in [2]. ERPD is now plotted on the y-axis and only trials with word-error-rate (WER) less than 10% were included in the analysis. Although 0% WER is expected in quiet conditions, with synthetic speech of poor quality this isn't feasible: too much data would be discarded under such a strict criterion. The improvements in analysis explained in the previous section

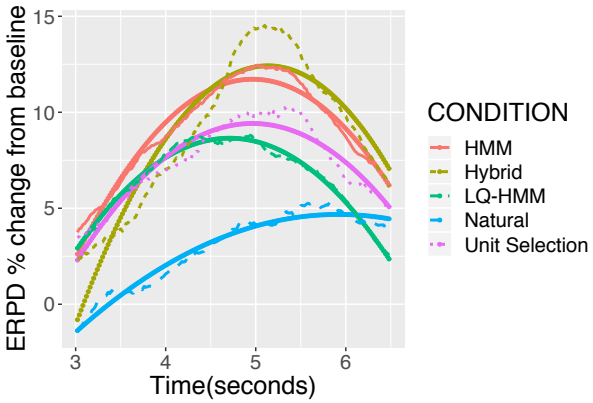

Figure 1: *Quiet condition: ERPD % change from the baseline across all participants and conditions. Dotted: Raw data, Solid: Quadratic Model fit*

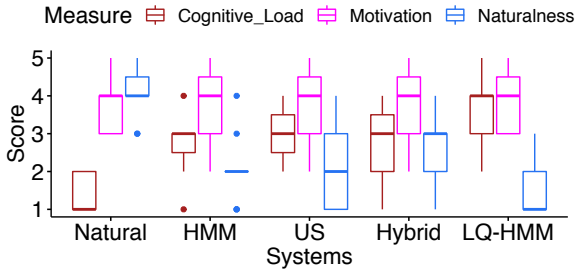

Figure 2: *Quiet condition: Self-reported measures*

were made[1]. Two participants were excluded from the analysis as more than 50% of their trials were discarded during pre-processing.

The raw data and GCA model fit are shown in Figure 1. The quadratic model provided a fairly good fit to the data. There was a significant effect of condition on the intercept, linear and quadratic terms. The intercept and linear term was significantly different for all conditions($p \leq 0.05$ for all comparisons). Significant differences in the quadratic term were found only for the Natural and Hybrid condition. The estimates are shown in Table 2. Hybrid has a significantly sharper peak than all other conditions. Natural is significantly flatter in peak in comparison to all other conditions. Self-reported measures are presented in Figure 2.

3.2. Experiment 2: Noise at -1dB SNR

In this experiment, trials with WER $\geq$ 20% were excluded. As mentioned earlier, the expected intelligibility level estimated from the psychometric curve in [10] was 80%. One participant was excluded from the analysis as more than 50% of their trials were discarded during pre-processing.

The raw data and GCA cubic model fit are shown in Figure 3. The cubic model fitted the data much better than the quadratic model and a significant improvement in all time terms were found when model comparison was performed. The cubic model fitted relatively well for all the synthetic speech condi-

[1]Whilst improving the analysis, we discovered a minor error in the previous analysis. This has now been fixed and thus Figure 1 in this paper replaces Figure 6 in [2]

Table 2: *Summary of estimates of quadratic term (time2) for all statistically significant conditions ($p \leq 0.05$) for the quiet condition. (Top) Baseline: Natural, (Bottom) Baseline: Natural*

Conditions	Estimate	Standard Error
time2:Natural	-8.93	5.57
time2:HMM	-20.68	2.17
time2:Hybrid	-25.56	2.14
time2:LQ-HMM	-20.07	2.24
time2:Unit Selection	-16.94	2.24
time2:Hybrid	-34.49	5.54
time2:HMM	4.88	2.12
time2:Natural	25.56	2.14
time2:LQ-HMM	5.48	2.18
time2:Unit Selection	8.62	2.18

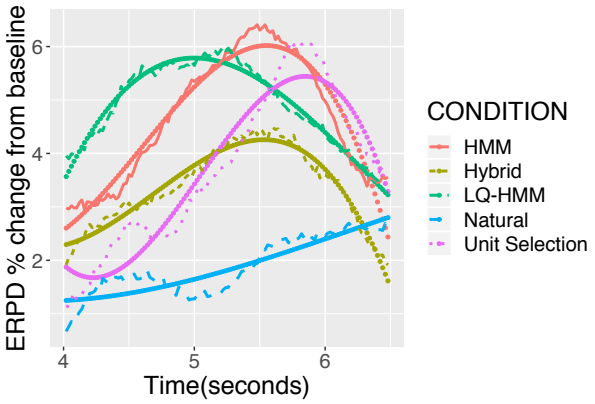

Figure 3: *-1dB SNR: ERPD % change from the baseline across all participants and conditions*

tions but is slightly under-fit for the Natural condition. Nevertheless, the natural condition had more of a linear shape and the behaviour in the response differs significantly to all other conditions. There was a significant effect of condition on the intercept term for almost all comparisons. LQ-HMM and US were the the only pair which did not differ to each other in intercept. In the linear term, all comparisons were statistically significant except the Hybrid and Natural pair. Hybrid and Natural have a less steep slope compared to all other conditions. In the quadratic term, Natural and HMM were the only conditions that were significantly different to all other conditions. In the cubic term, only LQ-HMM is statistically different to all other conditions. This is evident in the way the LQ-HMM returns to baseline. The quadratic estimates in Table 3 shows that HMM has the sharpest peak and Natural has the flattest. LQ-HMM, Hybrid and US are similar in peak shape. Self-reported measures are presented in Figure 4.

3.3. Experiment 3: Noise at -3dB SNR

In this experiment, trials with WER $\geq$ 40% were excluded to correspond with an intelligibility level of at least 60%. Two participants were excluded from the analysis as more than 50% of their trials were discarded during pre-processing.

The raw data and GCA cubic model fits are shown in Figure 5. The cubic model fitted the data much better than the quadratic model and a significant improvement in all time terms

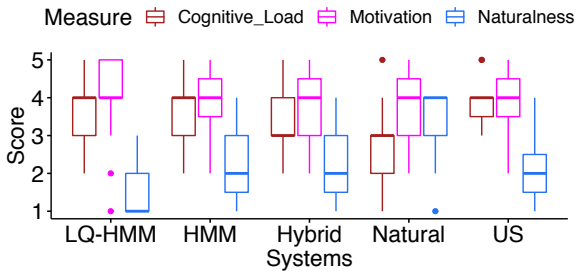

Figure 4: *-1dB SNR: Self-reported measures*

Table 3: *Summary of estimates of quadratic term (time2) for all statistically significant conditions ($p \leq 0.05$) for the -1dB condition. Baseline: HMM*

Conditions	Estimate	Standard Error
time2:HMM	-12.79	2.76
time2:Hybrid	4.73	1.84
time2:Natural	11.84	1.79
time2:LQ-HMM	5.04	1.76
time2:Unit Selection	5.27	1.93

were found when model comparison was performed. The cubic model fits the data almost perfectly. There was a significant effect of condition on the intercept term for almost all comparisons. LQ-HMM and Natural and LQ-HMM and US pairs did not differ in intercept. In the linear term, all comparisons were statistically significant except the Hybrid and HMM pair. This indicates these two conditions were equivalent in slope. In the quadratic term, all comparisons were statistically different to each other except the Hybrid and LQ-HMM pair. In the cubic term, only Hybrid and LQ-HMM pair were found to statistically different. The quadratic estimates in Table 4 show that US has the sharpest peak followed by HMM, LQ-HMM, Hybrid and then Natural with the flattest peak. This ordering suggests that as quality deteriorates, the flatter the peak becomes. Self-reported measures are presented in Figure 6.

4. Discussion

Listening to synthetic speech in quiet

Using GCA, Hybrid had the highest ERPD and the sharpest peak, whilst the natural condition had the lowest ERPD and flattest peak. This result for Natural is as expected: not much cognitive effort is exerted when listening to natural speech in quiet conditions [13]. An interesting observation is that the order of the peaks for the synthetic speech conditions show an inverse trend to what was reported in the self-report measures in Figure 2. Self-reported CL increases as quality decreases but peak pupil dilation appears to decrease. If cognitive effort is what we are measuring the reverse should be observed and the Hybrid system should be the lowest. Furthermore, the LQ-HMM condition was specifically selected due to its poor quality; since it did not induce the greatest pupil response, this raises a red flag. We firmly believe that listening effort is not being measured here, but rather attention. Pupillometry studies with degraded signals show that intelligibility declines with greater degradation, and greater loss of quality leads to increased pupil dilation [14, 15]. However, when high quality degraded speech is compared to

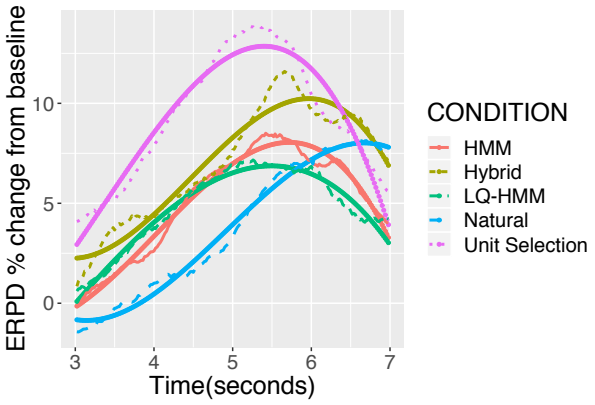

Figure 5: *-3dB SNR: ERPD % change from the baseline across all participants and conditions Dotted: Raw data, Solid: Cubic Model fit*

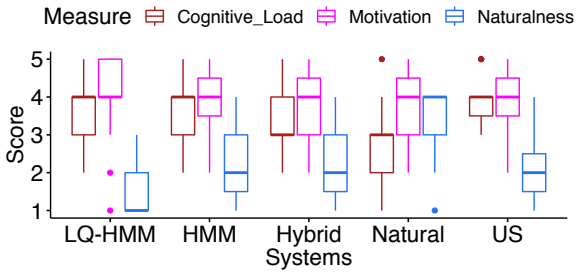

Figure 6: *-3dB SNR: Self-reported measures*

natural speech (i.e. intelligibility does not differ) then degraded signals elicit a relatively smaller pupil dilation. It appears that degraded signals appear to obscure acoustic cues, which engage attention in the processing of natural speech [12]. Similarly, in our work as the quality of synthetic speech increased, pupil dilation was smaller and thus listeners were probably more engaged. Additionally, the change in ERPD in the quiet condition in general exceeds the change in ERPD in the noise conditions as shown in Figure 7. We know for certain that listening in noise is harder than quiet, yet the ERPD is greater in quiet. Therefore, if cognitive effort isn't being measured, the only reasonable explanation is that engaged attention is measured. In quiet conditions, a greater pupil response is therefore more favourable.

Listening to synthetic speech in noise

In the easier SNR condition, the Natural condition has the lowest peak pupil response, in line with natural speech being less cognitively demanding than synthetic speech even in adverse conditions. In terms of slope, Hybrid and Natural reach peak pupil dilation with similar steepness. They both have the lowest ERPD and also have the lowest self-reported CL scores compared to all other conditions in Figure 4. Based on these results, Hybrid and Natural appear to impose the lowest load on the listener. Natural speech however differs to all synthetic conditions in peak shape, which is found to be the flattest. On the other hand, the HMM condition has the highest ERPD and has the sharpest peak according to the statistics. But, HMM, US and LQ-HMM are all perceived with similar self-reported load. Under adverse conditions, the poor quality systems become more difficult to separate. In the cubic term, LQ-HMM was the only condition to differ to all other conditions. It also scored the lowest naturalness score. Although it wasn't statistically different to the HMM and US conditions in peak, it differs in the manner in which it returns to the baseline.

Table 4: *Summary of estimates of quadratic term pairs which were statistically significant ($p \leq 0.05$) for the -3dB condition. Baseline: Unit Selection*

Conditions	Estimate	Standard Error
time2:Unit Selection	-35.68	6.63
time2:HMM	12.83	1.89
time2:Natural	33.66	1.89
time2:LQ-HMM	18.10	1.88
time2:Hybrid	20.35	1.93

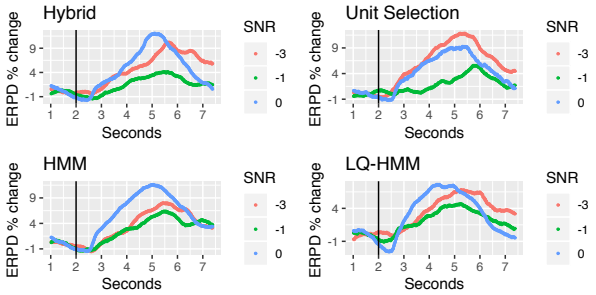

Figure 7: *ERPD % change from the baseline for each SNR level for condition*

In the harder SNR condition, the self-reported measures show that Natural, Hybrid and US have higher CL scores than HMM and LQ-HMM. It is interesting that the low quality systems are perceived to have a lower CL than the high quality systems. These scores also correspond with the peak pupil dilation, which shows that US, Hybrid and Natural have the higher ERPDs. The shape of the peaks, however, reflect the opposite. The flatter the peak, the higher the quality with the exception of the LQ-HMM condition found to be equivalent to the Hybrid system. Finally, for the poor quality systems: LQ-HMM and HMM, a small difference between the two SNRs were observed in Figure 7, indicating that ceiling in cognitive load was already reached at the higher SNR. This could indicate why the load was perceived to be less. For the high quality conditions: US, Hybrid and Natural larger differences were observed.

5. Conclusion

In our previous study, uncertainties regarding exactly what was being measured were raised. The current study attempted to resolve some of those. Our results show that, in quiet listening conditions, pupil dilation reflects engaged attention. In noisy conditions, increased pupil dilation for high quality synthetic speech indicates that listening effort increases as signal-to-noise ratio decreases whilst for low quality systems, ceiling is reached at easier SNR levels. Using GCA analysis both the slope and peak shape detected differences in listening effort between the various text-to-speech systems.

6. Acknowledgements

This project has received funding from the EU's H2020 research and innovation programme under the MSCA GA 675324.

7. References

[1] O. Simantiraki, M. Cooke, and S. King, "Impact of different speech types on listening effort," *Proc. Interspeech 2018*, pp. 2267–2271, 2018.

[2] A. Govender and S. King, "Using pupillometry to measure the cognitive load of synthetic speech," *Proc. Interspeech 2018*, pp. 2838–2842, 2018.

[3] E. H. Hess and J. M. Polt, "Pupil size in relation to mental activity during simple problem-solving," *Science*, vol. 143, no. 3611, pp. 1190–1192, 1964.

[4] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources." *Psychological bulletin*, vol. 91, no. 2, p. 276, 1982.

[5] A. A. Zekveld, S. E. Kramer, and J. M. Festen, "Pupil response as an indication of effortful listening: The influence of sentence intelligibility," *Ear and hearing*, vol. 31, no. 4, pp. 480–490, 2010.

[6] ——, "Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response," *Ear and hearing*, vol. 32, no. 4, pp. 498–510, 2011.

[7] 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.

[8] D. Mirman, *Growth curve analysis and visualization using R*. Chapman and Hall/CRC, 2017.

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

[10] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[11] M. B. Winn, D. Wendt, T. Koelewijn, and S. E. Kuchinsky, "Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started," *Trends in hearing*, vol. 22, pp. 1–32, 2018.

[12] A. E. Wagner, P. Toffanin, and D. Başkent, "The timing and effort of lexical access in natural and degraded speech," *Frontiers in Psychology*, vol. 7, p. 398, 2016.

[13] 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.

[14] A. A. Zekveld and S. E. Kramer, "Cognitive processing load across a wide range of listening conditions: Insights from pupillometry," *Psychophysiology*, vol. 51, no. 3, pp. 277–284, 2014.

[15] T. Koelewijn, A. A. Zekveld, J. M. Festen, and S. E. Kramer, "Pupil dilation uncovers extra listening effort in the presence of a single-talker masker," *Ear and Hearing*, vol. 33, no. 2, pp. 291–300, 2012.