Cognitive Load of Modern TTS Systems Under Noisy Conditions

Avashna Govender[1], Simon King[2]

[1]Council for Scientific and Industrial Research, South Africa
[2]Centre for Speech Technology Research, University of Edinburgh, United Kingdom

Abstract
Cognitive load of text-to-speech (TTS) synthesis systems measured in the past consistently showed that processing synthetic speech is more difficult to process than human speech in the presence of noise. However, the systems previously evaluated are no longer state-of-the-art. The quality produced by modern TTS systems are considered indistinguishable from human speech. Does this mean that the cognitive load demanded by such systems are now equivalent to that of human speech? The work presented in this paper, sets out to answer this question by measuring the cognitive load of modern TTS systems under noisy conditions. Results show that the gap of cognitive load demanded by TTS and human speech is reducing when listening to systems such as Tacotron 2 and Fastspeech 2. However, differences in cognitive load between these systems are still present. Therefore, despite modern TTS systems producing high quality speech, not all of them demand the same amount of cognitive load and thus not all TTS systems will provide the same user experience when embedded into real-world applications. Interestingly, results suggest that vocoded speech demands the same cognitive load as human speech which shows that it is possible to generate synthetic speech that can impose cognitive load that is equivalent to that of human speech.

Keywords
text-to-speech, pupillometry, cognitive load, listening effort

1. Introduction

Speech technology is increasingly becoming popular and therefore evaluating the users' experience is crucial. An important aspect of evaluating the users' experience is by understanding the difficulty experienced by the listener - if any - when listening to synthetic speech. To understand the difficulty of listening, one needs to understand how synthetic speech interacts with the human cognitive processing system whilst listening to it. In other words, a measurement of *cognitive load* is required [1]. Previous work has investigated the cognitive load of synthetic speech on various text-to-speech (TTS) systems in the past [2, 3, 4]. All of which have shown to demand a higher cognitive load than human speech. However, architectures in TTS are constantly evolving and the most recent TTS systems are capable of producing synthetic speech that is considered to be indistinguishable from human speech in terms of intelligibility and naturalness [5]. This therefore makes us question whether the cognitive load of modern TTS systems is also becoming indistinguishable from human speech. In this work, we set out to

Cognitive AI 2023, 13th-15th November, 2023, Bari, Italy.
agovender1@csir.co.za (A. Govender); simon.king@ed.ac.uk (S. King)

measure the cognitive load of modern TTS systems to determine whether synthetic speech produced by modern TTS systems are still more difficult to listen to than human speech.

2. Experimental design

2.1. Models evaluated

Two state-of-the-art TTS systems were selected for the evaluation, namely Tacotron 2 and Fastspeech 2 which is an adapted version of the original and is not publicly available. As a lower bound in the evaluation, an older model, the Merlin[1] TTS system was included. TTS systems comprise of two key components, namely the acoustic model and the vocoder. The acoustic model is responsible for the conversion from the text representation to an acoustic representation whilst the vocoder is responsible for generating the speech from the acoustic representation. To evaluate whether contributions to increased cognitive load stem from the acoustic model alone and not the vocoder, we included samples generated by the vocoder in the evaluation. All models were implemented in conjunction with the MultiBand-Melgan vocoder. As the upper bound, human speech taken from the original samples from the dataset were included.

2.2. Experimental setup

Our same pupillometry paradigm proposed in [3] was used to measure the cognitive load of the various models and human speech. The experiment was set-up in the same manner as reported in [3]. An SR-Eyelink eye tracker was used to measure the pupil response in a light and sound controlled lab whilst participants' listened to audio samples in the presence of noise through headphones. Three experiments were conducted. Each of them measuring the cognitive load of the various systems in the presence of speech-shaped noise at SNRs -1 dB (Exp. A, N=15), -3 dB(Exp. B, N=20) and -5 dB (Exp. C, N=25) respectively. As in [3], for each experiment, 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 scores were collected on a 5-point rating scale to support the results collected from the pupillometry paradigm. The same pre-processing and analysis procedures as reported in [4] were applied and the same event-related pupil dilation (ERPD) percentage formula was used. All analyses were carried out in R.

2.3. Analysis

In this work, Growth Curve Analysis (GCA) [6] was used to analyse the time course of the ERPD within a specific time period in which the peak is observed. The overall time course of the data was captured using a third-order (cubic) orthogonal polynomial with fixed effects of condition (various systems compared) and random effects of participant and item (sentence stimulus). Using GCA, parameter estimates are generated from the model fits and statistical

[1] This version is an adapted version of https://github.com/CSTR-Edinburgh/merlin

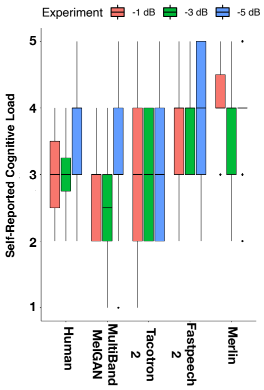

Figure 1: Boxplot of self-reported cognitive load in Exp. A (-1dB SNR), Exp. B (-3dB SNR) and Exp. C (-5dB SNR)

differences were obtained using post-hoc tests to get comparisons across all systems. In [3] a table is presented that describes what each time term represents. In this work, we focus only on the intercept term that represents the overall mean pupil dilation.

3. Results

3.1. Recall accuracy

After each trial, the listener was expected to repeat the sentence they heard verbatim. Recall accuracy (RA) was calculated by summing additions, substitutions and insertions and dividing by the total number of words in the sentence. Sentences consisted of approximately 8 words. Recall accuracy across all experiments are 75%, 64% and 51% respectively. The expected recall accuracy at -1dB, -3dB and -5dB as reported in [7] are approximately 80%, 60% and 40% for human speech respectively. Therefore, the recall accuracy for the experiments in this work were in line with the expected correct percentages except for the -1dB experiment which is a consequence of the Merlin TTS system performing poorly (RA=66%). In all experiments Merlin performed the worst and was significantly different ($p < 0.05$) to all other systems except Fastspeech 2. All other systems were found to be equivalent. Since all systems (except Merlin) were found to be equivalent, we can be certain that contributions to increased cognitive load observed in this work will not be a consequence of poor intelligiblity.

3.2. Self-reported measures

The self-reported cognitive load measures are presented in a boxplot in Figure 1. Merlin and Fastspeech 2 are perceived to be the hardest to listen to whilst vocoded speech was reported as the easiest to listen to with human speech closely following.

3.3. Pupil responses

The intercept parameter estimates from the GCA for each experiment are presented in Table 1. and the average pupil responses in each noise condition are presented in Figure 2.

Table 1
GCA intercept parameter estimates across each experiment (Estimates in bold reflect systems found equivalent)

System	**-1dB**	**-3dB**	**-5dB**
Human	**3.61**	**4.63**	**5.54**
MultiBand-Melgan	4.69	**4.55**	**5.46**
Tacotron 2	**3.53**	6.38	6.24
Fastspeech 2	7.27	8.63	5.83
Merlin	5.57	6.00	5.73

For all systems, the intercept increases or is equivalent between -1dB and -3dB, except Merlin which appears to be similar across all noise conditions. In -5dB, Fastspeech 2 has a smaller estimate compared to -3dB and this decline in pupil response is clearly visible in Figure 2. Tacotron 2 has remained more or less constant between -3dB and -5dB. This would mean that cognitive load is either decreasing or equivalent when listening in an increased noise condition - which is unlikely. In [8], it is reported that when an evoked pupil response is smaller than expected, this suggests the possibility of the listener withdrawing from the task. Since we observe a reduced pupil response for Fastspeech 2 in the -5dB condition this result suggests that listeners have withdrawn from the task. In other words, the listener has reached their ceiling cognitive capacity in attempting to process the speech, perhaps as a result of being too challenging to listen to. Similarly, the mean for Tacotron 2 remained the same for -3dB and -5dB. Again, it is unlikely, that in the most difficult SNR condition, the listener is experiencing similar loads. It is more plausible that in -5dB, a reduced pupil response is also being observed for Tacotron 2. Since both systems reached cognitive ceiling capacity in -5dB, by comparing their estimates in -3dB, we see that Fastspeech 2 is significantly greater than Tacotron 2. Therefore, Tacotron 2 demands less cognitive load than Fastspeech 2. Since the estimates for Merlin were all constant, perhaps, even in the easiest condition, Merlin was too challenging to listen to and therefore evoked a small pupil response throughout. Thus, from all systems evaluated, Merlin is the most difficult TTS system to listen to. This is not surprising as the recall accuracy for Merlin was poor, self-reported cognitive load was high as it was deliberately selected as the lower bound. For human speech, the pupil response increases gradually as the SNR decreases but is still manageable in -5dB. Vocoded speech is found to be equivalent to human speech in -3dB and -5dB but in -1dB it appears to behave differently to all other conditions and systems (see Figure 2). Overall, these findings suggest that vocoded speech is equivalent in cognitive load to human speech under noisy conditions. This finding is also an important one, as it shows that increased cognitive load contributions in TTS systems do not stem from the vocoder but rather from the acoustic model.

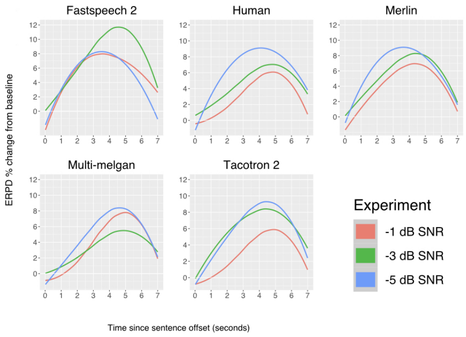

Figure 2: Time series line graph of cubic model fits for each system in Exp. A (-1dB SNR), Exp. B (-3dB SNR) and Exp. C (-5dB SNR)

4. Conclusion

The cognitive load of modern TTS was measured in this paper. From the self-reported cognitive load, vocoded speech was perceived to be the easiest to process with human speech closely following whilst Merlin was perceived to be the most difficult. Fastspeech 2 was perceived to be slightly more challenging to listen to than Tacotron 2. By evaluating the cognitive load of TTS systems using the pupillometry paradigm, the pupil response revealed the same results, thereby validating the self-reported measures. These findings suggests that modern TTS systems are becoming more manageable to listen to even in noisy conditions. More specifically, such results reveal that Merlin, a statistical parametric speech synthesis (SPSS) model [9] demands high cognitive load and should not be used in real-world solutions that embed TTS. Fastspeech 2 [10] and Tacotron 2 [11], both utilise a sequence-to-sequence based architecture which is shown to reduce cognitive load compared to SPSS. However, since differences were still observed between these 2 state-of-the-art systems, a deeper dive is necessary to understand where exactly increased cognitive load contributions stem from. Furthermore, given that vocoded speech has demanded the least cognitive load, this finding shows that increased cognitive load in TTS systems does not stem from the vocoder. In conclusion, despite modern TTS systems producing high quality speech, there are still differences between them, vocoded and human speech in terms of their cognitive load. Modern TTS systems are therefore moving in the direction of being equivalent to human speech but not all systems will provide the same user experience. Therefore the listeners' experience will vary depending on the architecture that is used within a given application. Given that we have identified that Fastspeech 2 versus Tacotron 2 imposes differing cognitive loads in this work, this information becomes valuable as it allows us to further unpack where possible contributions to increased cognitive load within the model stem from. Such information informs us on how to develop better TTS models that are equivalent to human speech or if not, even better. Evaluating the cognitive load of TTS is therefore necessary for selecting the right architectures to be embedded into real-world applications such that the lowest possible strain on listeners using them is imposed.

Acknowledgments

I would like to acknowledge Ella Crocker for collecting the pupil responses for this work.

References

[1] R. McGarrigle, K. J. Munro, P. Dawes, A. J. Stewart, D. R. Moore, J. G. Barry, 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 53 (2014) 433–445.

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

[3] A. Govender, A. E. Wagner, S. King, Using pupil dilation to measure cognitive load when listening to text-to-speech in quiet and in noise., in: Interspeech, 2019, pp. 1551–1555.

[4] A. Govender, C. Valentini-Botinhao, S. King, Measuring the contribution to cognitive load of each predicted vocoder speech parameter in DNN-based speech synthesis, in: 10th ISCA Speech Synthesis Workshop, 2019, pp. 121–126.

[5] Y. Shiga, J. Ni, K. Tachibana, T. Okamoto, Text-to-speech synthesis, Speech-to-Speech Translation (2020) 39–52.

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

[7] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, Y. Tang, Evaluating the intelligibility benefit of speech modifications in known noise conditions, Speech Communication 55 (2013) 572–585.

[8] A. E. Wagner, P. Toffanin, D. Başkent, The timing and effort of lexical access in natural and degraded speech, Frontiers in Psychology 7 (2016) 398.

[9] H. Zen, A. Senior, 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.

[10] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu, Fastspeech 2: Fast and high-quality end-to-end text to speech, arXiv preprint arXiv:2006.04558 (2020).

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, in: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2018, pp. 4779–4783.