PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Evaluating Cognitive Load of Text-To-Speech (TTS) synthesis

Avashna Govender, Cassia Valentini-Botinhao, Simon King

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

Abstract

Current evaluation methods for text-to-speech (TTS) synthesis rely solely on subjective rating scores. These tests typically account mostly for how natural or intelligible the voice is. With state-of-the-art systems, these measures are approaching ceiling and therefore alternative measures such as the cognitive load may become more meaningful. To our knowledge, there is little or no recent work evaluating the cognitive load of state-of- the-art text-to-speech systems. We use pupillometry as a measure of cognitive load. The pupil has been found to dilate upon increased cognitive effort when carrying out a listening task. Currently we are evaluating speech generated by a Deep Neural Network TTS synthesiser. In our method, we generate stimuli that step incrementally from natural speech to synthesized speech by changing only a single feature at a time. Stimuli are presented to listeners in speech-shaped noise conditions.

Keywords: text-to-speech, evaluation, cognitive load

1 INTRODUCTION

Text-to-speech (TTS) is artificial speech that is generated using a computer when a transcript is provided as input and the corresponding speech waveform is the output. Many real-world applications today, like voice-assistants such as Alexa and Google Home, speak using TTS. As these applications become more popular, the impact that such technology has on the end-user becomes important. Some studies have lead us to believe that whilst listening to synthetic speech our human cognitive processing system is placed under greater demand than listening to human speech. Therefore, if the quality of synthetic speech is not as high as natural speech this could lead to negative implications such as fatigue. Current evaluation methods for text-to-speech synthesis is limited and often does not include any measurement of cognitive load. Cognitive load was last measured for TTS when rule-based systems existed and therefore have become outdated. Cognitive load has yet to be measured on state-of-the-art TTS systems. Therefore, the aim of this work is to measure cognitive load of Deep Neural Network (DNN)-based TTS and understand the shortfalls in comparison to natural speech that lead to an increased cognitive load.

2 METHODOLOGY

2.1 Cognitive load measurement

Initial studies in the 1960's [5] measured pupil size while observers made pitch judgments. Results showed that a substantial dilation occurs immediately after the presentation of the comparison tone. Also, the size of the response was found to be closely correlated to the difficulty of the discrimination task. These results provided support for using pupillometry as an index of cognitive load. Recently, pupillometry has become popular in measuring cognitive load in speech understanding. It is typically referred to as listening effort [7, 6, 11]. Findings suggest listening effort correlates (positively) with pupil dilation. The more effort utilized, the larger the pupil dilates. These studies investigated listening effort of natural speech in the presence of noise which is intuitively more effortful. Therefore, following a similar approach we developed a pupillometry paradigm to measure the cognitive load of synthetic speech that was presented in [2] and is used in this work.

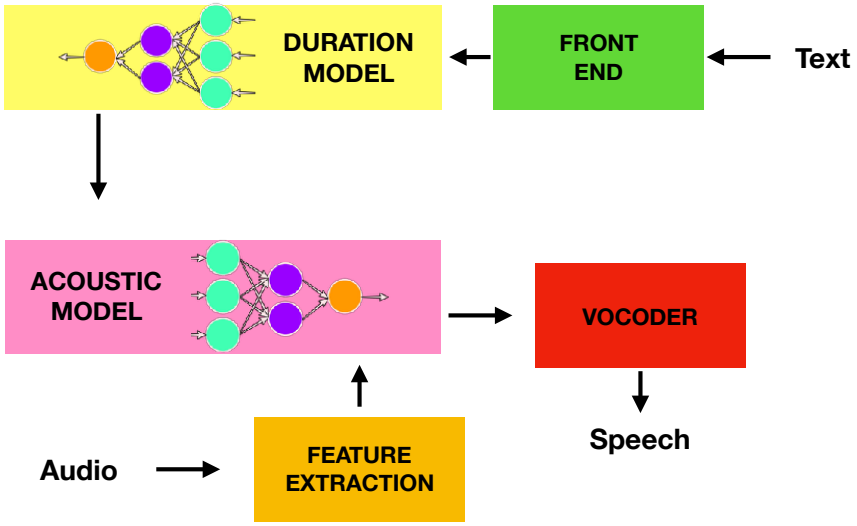

Figure 1. Architecture of a conventional DNN-speech synthesis system

2.2 Conventional DNN-speech synthesis

The architecture of a DNN-speech synthesis system is illustrated in Figure 2. In a conventional DNN system, text and audio pairs are used to train the models. Text is first processed using a *front-end* (eg., Festival in [1]). The objective of the front-end is to create a linguistic specification which captures important textual features like pronunciation, contextual features, stress patterns, part-of-speech etc. In the *feature extraction* block, the vocoder (eg., WORLD vocoder in [8]) extracts acoustic speech parameters from natural speech. Following the standard "build your own voice" recipe in Merlin [10], the features extracted are mel-cepstral coefficients (MCC), logarithmic fundamental frequency (F0) and band-aperiodicities (BAP). Before training the duration model, forced alignment is performed to get frame-by-frame time-aligned labels. These time-aligned labels are used to train the *duration model*. The *acoustic model* is then trained frame-by-frame using the linguistic features as input and the acoustic features as output. At synthesis time, duration predicted by the duration model are used by the acoustic model to generate the acoustic speech parameters features frame-by-frame. These parameters are then passed to the *vocoder* which the generates the waveform.

2.3 Experimental Conditions

The conditions evaluated consist of varying configurations of a DNN-based speech synthesis that steps gradually from natural to synthetic speech by changing only one acoustic speech parameter (described in Section 2.2) at a time. To do this, a full DNN TTS system was first trained. Then, to construct the intermediate configurations (A and B) we swapped the predicted mel-cepstral coefficients with the mel-cepstral coefficients extracted from natural speech. In the same manner we swapped the fundamental frequency speech parameters. Thereby creating mixtures of synthetic spectral features and "perfect" F0 features and vice versa. In another configuration (C), we generated acoustic features using natural duration as opposed to using the duration model. In addition, natural speech, vocoded speech and full TTS were included in the evaluation. In total 6 conditions were evaluated.

2.4 Experimental set-up

Our pupillometry experimental set-up comprises an SR-Eyelink eye-tracker that collects pupil measurements at 500Hz. Participants are recruited and the experiment lasts 30-45 minutes. The task of the listener is to fixate on

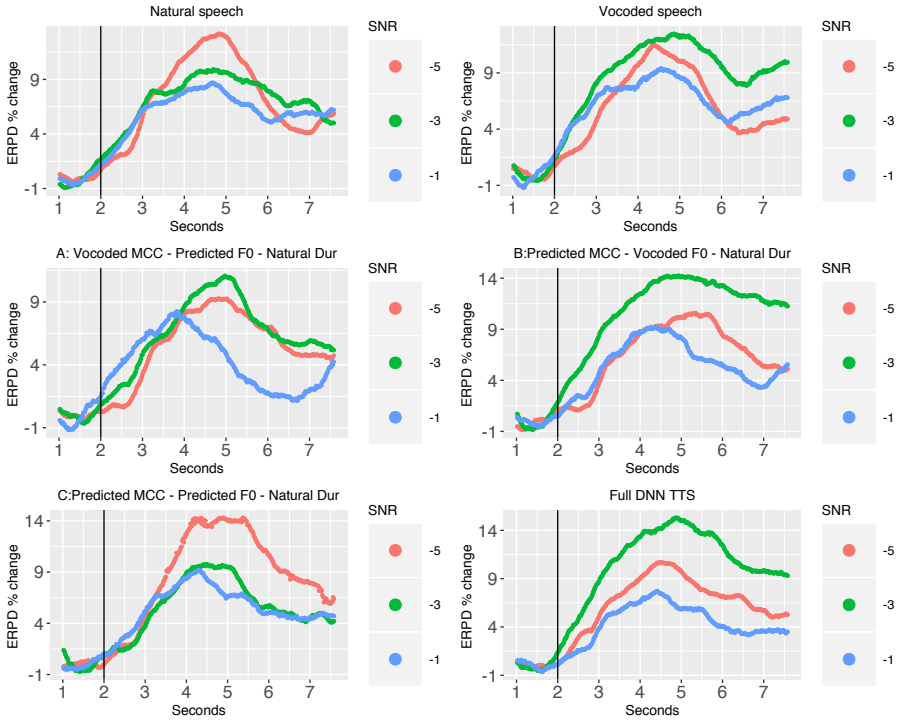

Figure 2. Pupil dilations when listening in -1dB, -3dB, -5dB, Dur: Duration)

a black cross shown in the centre of the screen whilst listening to sentences through headphones. To confirm the listener is listening and paying full attention they are expected to repeat the sentence. This is also used to calculate recall accuracy. The experiment is divided into blocks each containing one of the TTS system configurations evaluated. After each block, the listener is asked to rate their: difficulty in listening, perception of how natural the voice sounded and motivation to pay attention.

2.5 Stimuli

For each condition, 100 sentences from the Glasgow Herald Newspaper were synthesised. Each synthesised sentence was mixed with speech shaped noise at signal-to-noise-ratios of -1dB, -3dB and -5dB. 54 Native English participants were recruited and divided equally in each of the three experiments: -1dB, -3dB and -5dB.

3 RESULTS AND DISCUSSION

All pupil data collected was first processed prior to analysis which includes: Trial exclusion, deblinking and downsampling. Figure 2 presents the average of all remaining trials across all participants for each condition evaluated. Pupil dilation is presented as the event related pupil dilation (ERPD) % change from the baseline which is calculated as follows:

$$ERPD\% = (pupil\_size - baseline)\,\frac{100}{baseline} \tag{1}$$

where pupil_size is a single pupil size sample in the trial and the baseline is the average of all pupil size samples that falls in a 1 second window prior to the onset of the sentence (illustrated by the black line in Figure 2. This equation is applied to each sample from the onset on the sentence until the verbal response.

In all conditions we observe that the ERPD is the lowest when listening in noise at -1dB SNR. This is in line we what we expect, if the task is easy then pupil dilation should be low. For *Natural speech* and configuration *C* we observe that the ERPD remains the same even in the -3dB condition. For these two conditions, this suggests that the listening effort remained the same even though the SNR level was more challenging. Only in the most difficult SNR level at -5dB, an increase in pupil response was observed. For *Vocoded speech*, we observe that in both -3dB and -5dB they pupil response remains the same. This suggests that listeners found it equally effortful at -3dB and -5dB. The ERPD evoked at -3dB for *Vocoded speech* was the same height as that for -5dB in *Natural speech*. Therefore, this suggests that listeners reached ceiling for *Vocoded speech* already at the easier SNR of -3dB. For the remaining three configurations *A and B* and *Full DNN TTS*, we observe that at the -5dB SNR, the evoked pupil response was lower than -3dB. A possible explanation for this finding is that listeners could not cope when listening in -5dB SNR level. As consequence, the pupil response reflects fatigue. This result was also found in [9]. Two out of the three systems that struggle in the -5dB SNR comprise of spectral features that have been predicted from the text. Configuration *C*, which performed the best differs only to *Synthetic speech* in duration. Therefore, the key findings of this work is that poor spectral prediction and poor duration prediction contribute to an increased cognitive load.

(All pupil data was statistically analysed using growth curve analysis and presented in [3] together with a more detailed discussion of the findings.)

4 CONCLUSION

The work described here is one of many experiments that have been conducted towards measuring the cognitive load of text-to-speech synthesis [2, 4, 3]. The main contributions of cognitive load we have discovered thus far include poor spectral and duration prediction. Ongoing work aims to discover the influence of the vocoder itself by making comparisons with state-of-the-art phase and neural vocoders as opposed to the conventional source-filter vocoders used in this work.

5 ACKNOWLEDGEMENTS

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

References

[1] A. Black, P. Taylor, R. Caley, and R. Clark. The festival speech synthesis system, 1998.

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

[3] A. Govender, C. Valentini-Botinhao, and S. King. Measuring the contribution to cognitive load of each predicted vocoder speech parameter in dnn-based speech synthesis. *Submitted to Speech Synthesis Workshop (SSW) 2019*, 2019.

[4] A. Govender, A. E. Wagner, and S. King. Using pupil dilation to measure the cognitive load of synthetic speech in quiet and noise. *Submitted to Interspeech 2019*, 2019.

[5] D. Kahnemann and J. Beatty. Pupillary responses in a pitch-discrimination task. *Attention, Perception, & Psychophysics*, 2(3):101–105, 1967.

[6] 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*, 33(2):291–300, 2012.

[7] S. E. Kramer, A. Lorens, F. Coninx, A. A. Zekveld, A. Piotrowska, and H. Skarzynski. Processing load during listening: The influence of task characteristics on the pupil response. *Language and cognitive processes*, 28(4):426–442, 2013.

[8] M. Morise, F. Yokomori, and K. Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

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

[10] Z. Wu, O. Watts, and S. King. Merlin: An open source neural network speech synthesis system. In *SSW*, pages 202–207, 2016.

[11] A. A. Zekveld, S. E. Kramer, and J. M. Festen. Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. *Ear and hearing*, 32(4):498–510, 2011.