Interspeech 2018
2-6 September 2018, Hyderabad

Using pupillometry to measure the cognitive load of synthetic speech

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

It is common to evaluate synthetic speech using listening tests in which intelligibility is measured by asking listeners to transcribe the words heard, and naturalness is measured using Mean Opinion Scores. But, for real-world applications of synthetic speech, the effort (cognitive load) required to understand the synthetic speech may be a more appropriate measure. Cognitive load has been investigated in the past, when rule-based speech synthesizers were popular, but there is little or no recent work using state-of-the-art text-to-speech. Studies on the understanding of natural speech have shown that the pupil dilates when increased mental effort is exerted to perform a task. We use pupillometry to measure the cognitive load of synthetic speech submitted to two of the Blizzard Challenge evaluations. Our results show that pupil dilation is sensitive to the quality of synthetic speech. In all cases, synthetic speech imposes a higher cognitive load than natural speech. Pupillometry is therefore proposed as a sensitive measure that can be used to evaluate synthetic speech.

Index Terms: cognitive load, pupillometry, speech synthesis

1. Introduction

Developments in speech synthesis have provided dramatic improvements, from waveform concatenation [1] to Hidden Markov Models (HMM) [2] and most recently Deep Neural Networks (DNN) [3, 4]. The methods used to evaluate speech synthesis systems have however remained much the same. Evaluation is typically performed using subjective listening tests like those performed in the Blizzard Challenge [5, 6]. Intelligibility (as word error rate) and naturalness (as mean opinion score) are measured. As speech synthesizers attain near perfect intelligibility, they are increasingly being used in real-world applications. Therefore, the effort (cognitive load) required to understand the synthetic speech may be a more appropriate evaluation measure.

Pupillometry has become popular across many fields of research which include language processing, speech production and visual perception. Initial studies in the 1960s [7, for example] reported that the pupil dilates whilst solving arithmetic problems, and that the extent of the dilation was correlated with the difficulty of the problem. Since then, studies have consistently shown that there is a correlation between pupil dilation and the mental effort required to carry out a specific task [8, 9, 10, 11]. More recently, the pupil response has been used as an index of *listening effort* which is defined as the amount of mental effort allocated to a listening task [12, 13, 14, 15]. To our knowledge, ours is the first attempt to measure the cognitive load of *synthetic speech* using pupillometry. The work presented is a starting point for determining whether pupillometry is a viable measure.

2. Experimental Design

Three experiments were conducted to determine the sensitivity of the pupil response in evaluating synthetic speech. Experiment 1 investigated the effect of synthetic speech on listening effort. Experiment 2 examined whether results could be replicated across datasets. Experiment 3 investigated the effect of sentence type on listening effort. All experiments had a common set-up, and so we introduce this before discussing the three experiments. The set-up is a perceptual test using speech stimuli, which were sentences synthesised by one of 5 different speech synthesizers ('systems'), played to listeners whilst monitoring their pupil size using an eye tracker. One of the 5 systems was recorded natural speech.

2.1. Procedure

Structure Each experiment presented speech stimuli in 6 blocks. The first was a practice block. The remaining 5 blocks were constructed using a 5 x 5 Latin square design to ensure all listeners, systems and sentences were equally balanced. Block 1 had 5 trials, all using natural speech, to familiarise the listener with the experiment whilst avoiding exposure to the synthetic speech to be heard in the rest of the experiment. Each of the subsequent blocks had 20 trials with all speech stimuli taken from only one of the five systems. The sentences within each block were randomized.

Presentation Participants were tested individually in a light- and sound-controlled room. The eye tracker was calibrated for each listener at the start of the experiment using 9-point calibration. Listeners were instructed to focus on a fixation cross displayed at the centre of a computer screen for the duration of the trial. During each trial, an audio sample was played diotically to the listener through headphones. They were told that the cross would change from black to blue at the end of the trial, which is a signal for them to verbally repeat the sentence as accurately as possible. Their verbal responses were checked against the correct transcriptions to measure recall accuracy. In addition, subjective ratings were taken at the end of each block. Listeners were asked to rate the overall naturalness of the the audio samples (in the block they just listened to), the difficulty in listening to them and their motivation to pay attention, each on 5-point scales labelled *1 - unnatural* to *5 - natural*, *1 - very easy* to *5 - very difficult* and *1 - not motivated at all* to *5 - highly motivated* respectively. Listeners were allowed to take a break between blocks if desired. The experiment took approximately 30-45 minutes per listener.

Pupil size data collection Pupil data was collected at 500Hz using the SR Eyelink 1000 plus. The remote desktop mode was used which allowed the listener to move their head freely and not feel uncomfortable with a head mount. A target sticker was placed on the participant's forehead to aid the eyetracker with head movement.

10.21437/Interspeech.2018-1174

Pre-processing and analysis The mean and standard deviations (SD) of the pupil size, from 1 second before sentence onset until the start of the verbal response, were calculated. Pupil size values more than 2 SD smaller than the mean were coded as blinks. Trials for which more than 20% of the duration consisted of blinks were excluded. For the remaining trials, blinks were removed using linear interpolation. For ease of processing, the data was downsampled to 50Hz. Subsequently, baseline correction was performed on each trial by subtracting the mean baseline (1 second before sentence onset) from all the points across the trial. Typically, in experiments using only natural speech, incorrect recall of sentences is taken as an indication of loss of attention, and such trials would be excluded. Since synthetic speech was used in this experiment, partially-incorrect recall is more likely to be caused by poor quality stimuli. Therefore, only sentences with a high word error rate (WER $>$ 40%) were excluded. In addition, the first three sentences in each block were excluded from further analysis. A 5-point moving average filter was then applied to smooth each trial and remove artefacts. Analysis was performed by calculating the average pupil size, peak dilation and peak latency per participant per system. Peak dilation was defined as the highest value in the trial (before end of sentence). Peak latency was defined as the time of the peak relative to sentence onset. Mean dilation was calculated as the average across the trial. 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, a paired t-test with Bonferroni correction was used to determine which pairs of systems had statistically significant differences.

2.2. Participants

45 participants were recruited from university students and staff, ranging in age from 19 to 37 years, and divided across the 3 experiments (15 participants each). All participants were native English speakers; no hearing problems were self-reported.

3. Experiment 1

3.1. Speech Material

Stimuli presented to the participants were sentences generated by four synthesizers taken from the 2011[1] Blizzard Challenge [16] and natural versions from the same speaker. The synthesizers include: Hybrid, Unit Selection, Hidden Markov Model (HMM) and Low-Quality HMM system. All systems were created using approximately 16.6 hours of speech from a US English female professional speaker[2].

Since listening to natural speech is considered to be effortless in ideal (quiet) conditions [17], we were concerned that the pupil may not give a measurably large response due to insufficient cognitive load on the listener. In [8], semantically unpredictable sentences (SUS) were reported to evoke a greater pupil response than simple sentences. This motivated the use of SUS in this experiment.

3.2. Results

Recall accuracy is presented in Figure 1, and was at least 95% for all systems. There are no differences between the high quality speech synthesizers and natural speech. The Low-Quality HMM system was statistically significantly different to the rest of the systems, as expected. (ANOVA for repeated measures, F(4,56) = 5.3 with p=0.004 (p$\leq$0.05), n^2 = 0.2).

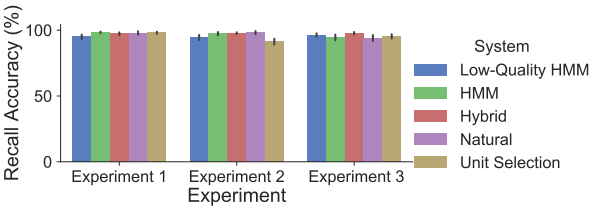

Figure 1: *Recall accuracy across all 3 experiments*

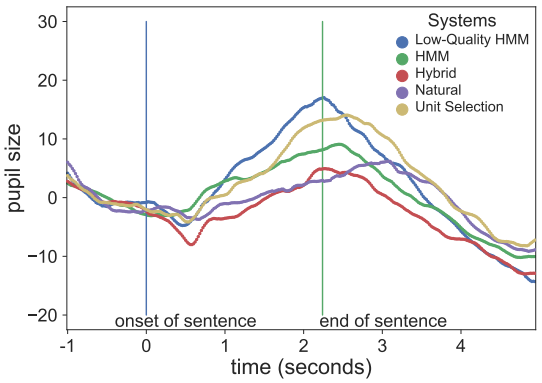

Figure 2: *Experiment 1: Pupil response for 4 speech synthesizers from the Blizzard Challenge 2011 and the corresponding natural speech for semantically unpredictable sentences*

The average pupil size across participants is presented in Figure 2. The differences in mean and peak pupil dilations between all systems were statistically insignificant. (ANOVA for repeated measures, Mean: F(4,56) = 1.5 with p=0.2 (p$>$0.05), n^2=0.04, Peak: F(4,56) = 2.09 with p=0.09 (p$>$0.05), n^2=0.09). The only statistical difference found was in the peak latency between the Natural speech and Low-Quality HMM system (ANOVA for repeated measures, F(4,56) = 4.21, with p=0.005 (p$\leq$0.05), n^2=0.16). For Natural, pupil dilation peaks at 3.1 seconds after sentence onset whereas for Low-Quality HMM it peaks much faster, at 2.3 seconds. This suggests that peak latency – a measure of listening effort in this experiment – could be correlated to either intelligibility or naturalness (discussed next) or both. At this point it is difficult to be sure which property of the speech led to faster pupil dilation. The Low-Quality HMM system is particularly poor in both naturalness and intelligibility which together could demand high effort from the listener. In Figure 2, the peak dilation correlates (negatively) more with intelligibility as measured in the Blizzard Challenge 2011 [16].

The self-reported measures for this experiment are presented in Figure 3 and show that the subjectively easiest system to listen to was Natural, followed by Hybrid and Unit Selection. Listeners had mixed ratings for HMM and found Low-Quality HMM most difficult. The naturalness ratings followed the same trend as the difficulty ratings, with HMM rated unnatural. The naturalness ratings in this experiment are worse than the findings of the Blizzard Challenge 2011[16]. Listeners motivation to pay attention was sustained across all systems.

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

[2] Freely available from `https://www.synsig.org/index.php/Blizzard_Challenge`

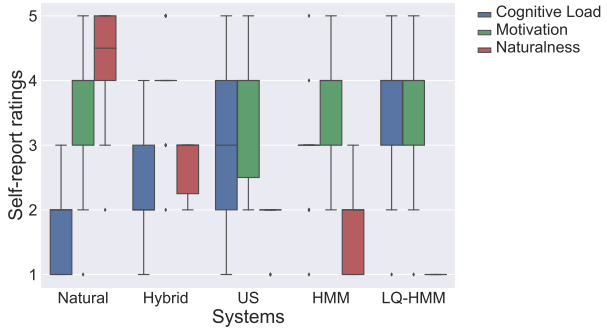

Figure 3: *Self-reported measures for Experiment 1*

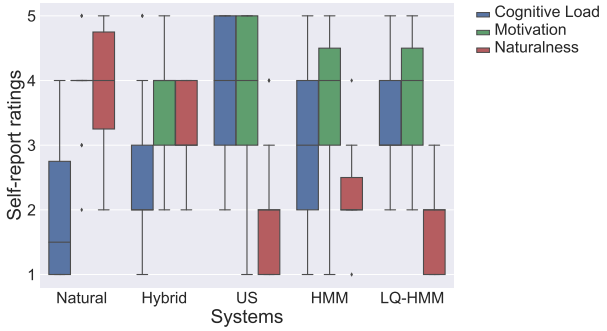

Figure 5: *Self-reported measures for Experiment 2*

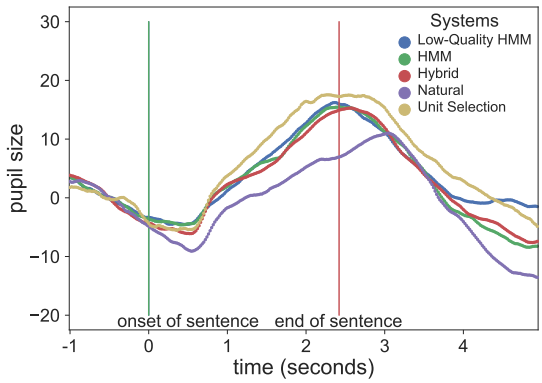

Figure 4: *Experiment 2: Pupil response for 4 speech synthesizers from Blizzard Challenge 2010 and the corresponding natural speech for semantically unpredictable sentences*

4. Experiment 2

4.1. Speech Material

This experiment replicates Experiment 1 with different speech material: stimuli presented to the participants were SUS from 5 systems: 4 synthesizers from the 2010 Blizzard Challenge [5] plus natural speech from the same speaker. The synthesizers were: Hybrid, Unit Selection, Hidden Markov Model (HMM) and Low-Quality HMM. All synthesizers were created using the 4 hour 'rjs' corpus of speech from a professional 50-year old male speaker with an RP accent.

4.2. Results

The recall accuracy is presented in Figure 1 and averaged 91% across systems. No differences between Natural, Hybrid and HMM systems were found. The Unit Selection system was statistically significantly different from all systems except the Low-Quality HMM (ANOVA for repeated measures, F(4,56) = 10.97 with p=1.2e-06 ($p \leq 0.05$), n^2 = 0.34). This is consistent with the relatively high WERs reported in [5]. The recall accuracy for Low-Quality HMM was statistically significantly different to Natural.

Differences in mean pupil size, peak pupil dilation and peak latency were found to be statistically insignificant between all systems. (ANOVA for repeated measures, Mean: F(4,56) = 1.8 with p=0.14 ($p>0.05$), n^2=0.03, Peak: F(4,56) = 0.9 with p=0.44 ($p>0.05$), n^2=0.02, Latency: F(4,56) = 1.44, p=0.23 ($p>0.05$), n^2=0.05). Although differences in peak latency did not reach significance, the response to Natural speech appears to be delayed in comparison to synthetic speech (Figure 4), as also observed in Experiment 1. The Hybrid, HMM and Low-Quality HMM peaks are bunched together with the Unit Selection system slightly separated from these. This result is consistent with the findings in Experiment 1, again suggesting that listening effort increases with decreasing intelligibility.

The self-reported measures for Experiment 2 are presented in Figure 5. Listeners found Natural the easiest to listen to, followed by Hybrid, HMM and Low-Quality HMM. Listeners found Unit Selection the most difficult. The naturalness ratings follow a similar trend to the difficulty ratings, with no differences reported between Unit Selection and Low-Quality HMM. The naturalness ratings in this experiment differ from the original findings in the Blizzard Challenge 2010 [5] where the Unit Selection system was rated more natural than HMM and Low-Quality HMM. This suggests that listeners' naturalness ratings in our experiment were biased by their understanding of the speech. Motivation to pay attention was sustained across the speech synthesizers but mixed ratings were received for the natural speech.

5. Experiment 3

5.1. Speech Material

Semantically Unpredictable Sentences are widely employed in intelligibility testing of synthetic speech, as a way to measure segmental intelligibility with reduced interference from the listener's strong predictive model of word sequences (and in doing so, to avoid a ceiling effect). But SUS have low ecological validity, especially for measuring 'real world' application performance [18, sec. 17.2.2]. Our interest in measuring cognitive load is motivated directly by the use of synthetic speech in real applications. Therefore, in this final experiment, the same systems as Experiment 1 were used, but the listeners were presented with meaningful sentences[3].

5.2. Results

Differences in recall accuracy in Figure 1 across all systems in Experiment 3 were again insignificant (ANOVA for repeated measures, F(4,56) = 0.7 with p=0.55 ($p>0.05$), n^2 = 0.04): all systems had at least 94% recall accuracy. The Low-Quality HMM system performed as well as all other systems, in contrast

[3]News sentences taken from Glasgow Herald newspaper used for the 2011 Blizzard Challenge

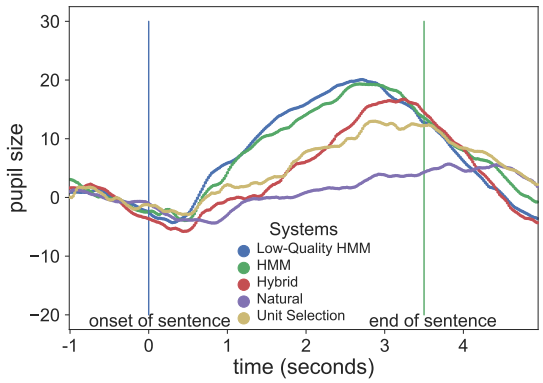

Figure 6: *Experiment 3: Pupil response for 4 speech synthesizers from Blizzard Challenge 2011 and the corresponding natural speech for meaningful sentences*

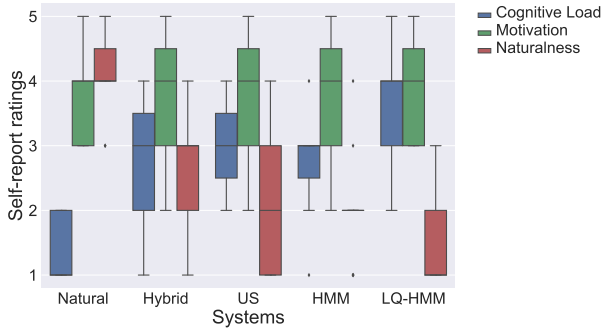

Figure 7: *Self-reported measures for Experiment 3*

to Experiment 1. This is of course as expected: listeners are better at recalling meaningful sentences than SUS, especially when quality is poor.

Differences in mean pupil size, peak pupil dilation and peak latency between all systems were found to be statistically insignificant (ANOVA for repeated measures, Mean: F(4,56) = 1.19 with p=0.32 (p>0.05), n^2=0.03, Peak: F(4,56) = 1.67 with p=0.17 (p>0.05), n^2=0.04, Latency: F(4,56) = 0.6, p=0.65 (p>0.05), n^2=0.02).

The peak pupil dilations for the synthesizers appear to be somewhat correlated to naturalness. Unit Selection generates waveforms by concatenating units from natural speech, which could explain the low peak. On the other hand, the HMM systems generate waveforms using a vocoder, and they yield higher peak pupil responses. The Hybrid system is a combination of Unit selection and SPSS and yields a peak that lies in the middle.

In Figure 6 it is clear that natural speech barely causes any pupil dilation at all. This is not surprising, and simply confirms that listening to meaningful natural sentences in quiet conditions is effortless [17]. In Experiment 1, which used speech from the same speaker, the SUS material did cause the pupil to dilate, indicating increased listening effort.

The peak latencies of the speech synthesizers are delayed in comparison to Experiment 1. We come back to this in the discussion.

The self-reported measures for this experiment are presented in Figure 7. Listeners found natural speech the easiest to listen to, and all the speech synthesizers were given similar ratings. Listeners also found the natural speech most natural followed by Hybrid and Unit Selection. Listeners gave mixed ratings for HMM and found Low-Quality HMM the most unnatural. Motivation to pay attention was sustained across the speech synthesizers but mixed ratings were received for the natural voice.

6. Discussion and Conclusions

Effect of synthetic speech on listening effort
Our results show that pupil dilation is sensitive to the quality of synthetic speech relative to natural speech. Differences in peak pupil dilation and peak latency between natural and synthetic speech were found. In all cases, synthetic speech imposes a higher cognitive load than natural speech. On the other hand, differences between speech synthesizers were more difficult to detect. We found a tendency towards differences, but these were not strong enough to reach significance. Alternative statistical tests, such as growth curve analysis, might have given more accurate estimates of significance.

Replication across datasets
Results in Experiment 1 and 2 were similar in relation to listening effort, which is negatively correlated with intelligibility. Results differed in absolute terms, presumably because intelligibility varied between the different Blizzard Challenge datasets (WER was lower for 2010 than 2011). But results were replicated across different datasets (and with different listeners).

Semantically unpredictable vs. Meaningful sentences
Changing from semantically unpredictable to meaningful sentences had an effect on peak latencies for synthetic speech. (No peak was found for listening to naturally-spoken meaningful sentences, since this is effortless.) Listening effort appears to be negatively correlated with intelligibility in Experiment 1 and 2 (using SUS), then negatively correlated with naturalness in Experiment 3 (using meaningful sentences). Using meaningful sentences is more ecologically valid, and appears to provide more sensitivity to naturalness. The set-up of Experiment 3 gives better insights into listening effort and naturalness, both of which are important for advancing the state-of-the-art.

Speech rate
One experimental variable that was not controlled in our experiments was speech rate. Although all synthesizers in any given year of the Blizzard Challenge are built from the same data, they may still generate speech at a different speaking rate. This is likely to affect absolute values of peak latency. Future experiments should take this into account.

7. 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`).

8. References

[1] 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.

[2] 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.

[3] 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*. Vancouver, Canada: IEEE, 2013, pp. 7962–7966.

[4] 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.

[5] S. King and V. Karaiskos, "The Blizzard Challenge 2010," Makuhari, Chiba, Japan, 2010.

[6] S. King, L. Wihlborg, and W. Guo, "The Blizzard Challenge 2017," Stockholm,Sweden, 2017.

[7] 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.

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

[9] J. Beatty and D. Kahneman, "Pupillary changes in two memory tasks," *Psychonomic Science*, vol. 5, no. 10, pp. 371–372, 1966.

[10] D. Kahnemann and J. Beatty, "Pupillary responses in a pitch-discrimination task," *Attention, Perception, & Psychophysics*, vol. 2, no. 3, pp. 101–105, 1967.

[11] D. Kahneman and J. Beatty, "Pupil diameter and load on memory," *Science*, vol. 154, no. 3756, pp. 1583–1585, 1966.

[12] 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.

[13] 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*, vol. 32, no. 4, pp. 498–510, 2011.

[14] ——, "Pupil response as an indication of effortful listening: The influence of sentence intelligibility," *Ear and hearing*, vol. 31, no. 4, pp. 480–490, 2010.

[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.

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

[17] 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.

[18] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.