Measuring the contribution to cognitive load of each predicted vocoder speech parameter in DNN-based speech synthesis

Measuring the contribution to cognitive load of each predicted vocoder speech parameter in DNN-based speech synthesis

Avashna Govender, Cassia Valentini-Botinhao, Simon King

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

a.govender@sms.ed.ac.uk, cvbotinh@inf.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

Listening to even high quality text-to-speech - such as that generated by a Deep Neural Network (DNN) driving a vocoder - still requires greater cognitive effort than natural speech, under noisy conditions. Vocoding itself, plus errors in predictions of the vocoder speech parameters by the DNN model are assumed to be responsible. To better understand the contribution of each parameter, we construct a range of systems that vary from copy-synthesis (i.e., vocoding) to full text-to-speech generated using a Deep Neural Network system. Each system combines some speech parameters (e.g., spectral envelope) from copy-synthesis with other speech parameters (e.g., F0) predicted from text. Cognitive load was measured using a pupillometry paradigm described in our previous work. Our results reveal the differing contributions that each predicted speech parameter makes to increasing cognitive load.

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

1. Introduction

Evaluation methods generally fail to consider the *cognitive load* imposed by listening to synthetic speech. This is especially concerning as synthetic speech demands the greatest effort even at favourable signal-to-noise ratios [1]. This highlights the negative impact synthetic speech potentially has on the human cognitive processing system. Therefore, a better understanding of how the cognitive load affects listeners is important to eventually suggest new ways of generating synthetic speech that demands low cognitive load.

Many perception studies have shown correlations between fluctuations in pupil size, measured using pupillometry, and changes in mental task load [2, 3, 4, 5]. These fluctuations could be related to changes in attention, stress, and working memory [6]. In speech understanding studies, the pupil response has often been used as an index for *listening effort*, i.e. the amount of mental effort allocated to a listening task [7, 8, 9, 10].

In our previous works, we used pupillometry to measure the cognitive load of synthetic speech [11, 12]. This method has proven to be reliable in showing pupil dilation as an index of listening effort under *noisy conditions*. In quiet conditions, however, we confirmed pupil dilation reflects more the engaged attention of the listener rather than mental effort [12]. Under noisy conditions, increased pupil dilation for high quality synthetic speech indicated that listening effort increases as signal-to-noise ratio decreases. For low quality systems such as HMM-based speech synthesis ceiling listening effort appears to be reached already at easier SNR levels. An HMM system was used in [1] which explains the great effort demanded compared to all other speech types. Recently, HMM-based speech synthesis has been, in most cases, replaced by Deep Neural Networks (DNNs). Therefore, in this work we investigate whether DNNs still demand greater cognitive effort than natural speech by gradually stepping from natural speech to a full DNN TTS system.

Table 1: *Summary of all configurations evaluated. MCC: mel-cepstral coefficients. BAP: band aperiodicities. Nat: natural. Pred: predicted. Voc: vocoded. System B is copy synthesis and System F is full text-to-speech.*

System	MCC	F0	BAP	DURATION
A	Human speech			
B	Voc	Voc	Voc	Nat
C	Voc	Pred	Pred	Nat
D	Pred	Voc	Pred	Nat
E	Pred	Pred	Pred	Nat
F	Pred	Pred	Pred	Pred

2. Experimental Design

We measured the cognitive load induced in listeners when listening to various types of speech ranging from natural speech to full text-to-speech.

2.1. Data and Implementation

A database consisting of speech sampled at 16 kHz from a British male speaker was used to train the DNN-based synthesis system. A total of 2072, 200 and 270 training, validation and testing sentences were used. Since we wish to measure cognitive load of synthetic speech for real applications, structurally correct and meaningful sentences for testing were used in all experiments, taken from the Glasgow Herald newspaper. All features were extracted using the WORLD vocoder [13]. 60-dimensional mel-cepstral coefficients (MCCs), 25 band aperiodicities (BAPs) and logarithmic fundamental frequency (F0) at 5 ms frame intervals were extracted. A DNN was trained using the Merlin toolkit [14] following the standard "build your own voice" recipe: acoustic and duration models each comprising 6 feed-forward hidden layers; each hidden layer has 1024 hyperbolic tangent units. Table 1 shows the configurations of systems constructed that vary from copy-synthesis (System B) to a full DNN text-to-speech system (System F). Systems C and D combine spectral parameters from copy-synthesis with F0 parameters predicted from text and vice versa. System E combines predicted spectral and F0 features at durations copied from natural speech recordings by forced alignment.

2.2. Experimental set-up

As in [11], the speech stimuli were played to listeners through headphones in a noise- and light-controlled room. Simultane-

ously, pupil size was measured using an eye tracker. All stimuli were mixed with speech-shaped noise at signal-to-noise ratios (SNRs) -1, -3 and -5 dB, chosen such that the cognitive load is increased whilst intelligibility remains close to ceiling. In accordance with the estimated psychometric function in [15] which related keyword scores to SNR for speech-shaped noise, the expected keyword correct percentages at -1, -3 and -5 dB are approximately 80, 60 and 45% for natural speech respectively. The procedure in this work was followed exactly in terms of structure, presentation and data collection as described in [12].

Similarly to [12], stimuli were blocked by system, resulting in 6 blocks, each containing 20 sentences. The block order was balanced using a 6×6 Latin square design to ensure all listeners, systems and sentences were equally represented and that no listener heard the same text more than once. At the end of each block, self-reported cognitive load (CL), motivation to listen, and naturalness scores were collected on 5-point rating scales (1 - very unnatural, very difficult and unmotivated;
5 - very natural, very easy and highly motivated)

2.3. Participants

72 Native English speakers with no self-reported hearing problems, aged 19 to 37 years, were recruited. The participants were equally divided between four experiments in which listening in quiet and each SNR condition was evaluated separately.

2.4. Pre-processing of pupil measurements

All pre-processing was the same as [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 300 ms, 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. The data was then downsampled to 50 Hz for faster processing. Subsequently, the Event Related Pupil Dilation (ERPD) percentage was computed. This was calculated using the equation in [16]. Some problems with the eye-tracker were experienced during data collection where a warning of the pupil size was shown. To ensure only viable responses were taken into account, a filter was applied to remove all trials where the ERPD was less than zero for more than 80% of the individual trial.

2.5. Pre-analysis

Experiment 1 (Quiet): Two participants were excluded from the analysis because more than 60% of their trials were discarded during pre-processing. The threshold was increased to 60% in this work compared to 50% in our previous work. At 50% exclusion, too many (14) participants would be excluded. Trial exclusion depends on blinking, spikes in pupil response and whether the eye-tracker lost the eye. Thus, the amount of data that can become unusable will vary from experiment to experiment. All trials with word-error-rate (WER) less than 10% were included in the analysis. If a threshold of 0% WER were used, too much data would be discarded. All retained trials averaged by system had a WER of less than 1%. In other words: intelligibility was at ceiling.

Experiment 2 (-1dB SNR): Two participants were excluded for the same reason mentioned above. Trials with WER $\geq$ 20% were excluded. In [15], the expected intelligibility level estimated from the psychometric curve was 80%. All retained trials averaged by system had a WER $\leq$ 2.5%. At this SNR level, listening to synthetic speech produced by a DNN-speech synthesis system was still close to ceiling. In [12], we found that at -1 dB SNR, intelligibility already started to suffer for those systems compared (ie. Unit Selection, HMM-based synthesis, and Hybrid speech synthesis). Therefore, the increased quality produced when using a simple feed-forward DNN architecture improves intelligibility under noisy conditions.

Table 2: *Summary of interpretation of each time term in GCA Formula: ERPD* $\sim (time1 + time2) * SYSTEM + (time1 + time2|SUBJECT) + (time1 + time2|ITEM) + (time1 + time2|GROUP)$

Term	**Interpretation**
Intercept	Overall mean pupil dilation
Linear (time1)	Overall rate of pupil dilation
Quadratic (time2)	Shape of peak

Experiment 3 (-3dB SNR): Two participants were excluded from the analysis for the same reason above. Trials with WER $\geq$ 40% were excluded to correspond with an intelligibility level of at least 60%. At this SNR, intelligibility starts to suffer. All retained trials averaged by system had a WER $\leq$ 10%.

Experiment 4 (-5dB SNR): One participant was excluded from this analysis as too much data was discarded. In this experiment, trials with WER $\geq$ 60% were excluded to correspond with an intelligibility level of at least 40%. This is slightly more lenient than the expected 45% level of intelligibility for natural speech calculated in [15]. All retained trials averaged by system had a WER $\leq$ 20%.

2.6. Analysis

In many listening effort experiments that use the pupillometry paradigm [8, 10, for example], eye-tracking data is analyzed using only the mean and maximum pupil dilation. These values are extracted by time-binning the pupil data and then selecting a single time window. This approach eliminates a lot of the meaningful information offered in eye-tracking data, notably changes over time are lost. In contrast, Growth Curve Analysis (GCA) [17] addresses these limitations and has become popular for the analysis of such data [18]. Using GCA, the time course of the ERPD within a specific time period in which the peak was observed was modelled using a second-order (quadratic) polynomial that makes the individual time terms independent. Using model comparisons we found no significant difference in the model fit beyond the second-order polynomial. A fixed effect of system (various configurations) and random effects of subject, group (with respect to the Latin square design) and item (sentence stimulus) were used on all time terms. Post-hoc tests were performed by changing the baseline condition and cycling through each of the six systems to get comparisons across all systems for each time term. Table 2 summarizes what each time term represents. Statistical significance (p-values) for individual parameter estimates were assessed using the normal approximation. All analyses were carried out in R.

3. Results

3.1. Pupil Responses (ERPD %)

The raw data and GCA model fits for each of the four experiments (when listening in Quiet, -1, -3, -5 dB SNR levels) are shown in Figure 1. In all cases, the quadratic model provided a fairly accurate fit to the data and significant improvement in all time terms were found during model comparisons ($p \leq 0.01$). Due to the large number of comparisons and to make the results more readable, we present in Table 3 the systems that had the highest and lowest parameter estimates. In the intercept term this corresponds with the highest and lowest mean pupil dilation, in the linear term, the steepest and flattest slopes and in the quadratic term, the sharpest and broadest peak shapes.

Table 3: *Summary of systems that evoked the (a) highest and lowest mean pupil dilation, (b) steepest and flattest slope, and (c) sharpest and broadest peaks. Multiple systems are shown when systems were found to be equivalent.*

(a)

Listening condition	Highest	Lowest
Quiet	C	E,F
-1 dB	A, B, F	C, D
-3 dB	D, F	C
-5 dB	E	C

(b)

Listening condition	Steepest	Flattest
Quiet	C, D, E	A, B
-1 dB	F	A
-3 dB	D,F	A, B, C
-5 dB	D, E	C, F

(c)

Listening condition	Sharpest	Broadest
Quiet	A, C, D, E	B, F
-1 dB	D, E	A, B, C, F
-3 dB	B, C, E, F	A, D
-5 dB	A,B,E	C, D, F

3.2. Self-reported measures

The self-reported measures collected for each listening condition is presented in Figure 2. The changes as we move from listening in quiet to listening in adverse noise conditions for each system compared are discussed below.

System A (human), maintains a median of 4 in terms of naturalness across all listening conditions. In quiet, CL is reported as 1 which is very easy to listen to. When the noise level increased, the CL also increased. However, CL stayed the same for the -3 and -5 dB conditions.

System B (vocoded), maintains a median of 4 in terms of naturalness across most listening conditions except -3 dB. In quiet, CL is reported as 2 which is easy to listen to. When listening in noise, CL increased as expected. In -1 dB and -3 dB the CL stayed the same. It is surprising that in the -5 dB condition, naturalness is perceived to be higher than in the -3 dB condition.

System C (vocoded MCC and predicted F0), maintains a median of 4 in terms of naturalness across most listening conditions except in quiet. Although the speech of this system was perceived to be less natural, listeners still reported CL as 2. As the noise level increased, the CL also increased. In the -5 dB condition, the cognitive load is reported as less difficult than the -3 dB condition.

System D (vocoded F0 and predicted MCC), has a median of 3 in terms of naturalness across most listening conditions. This score is lower than Systems A, B and C. Similar to System B in the -5 dB condition, naturalness is perceived to be the highest with a median of 4. Like System C, CL was reported as 2 in quiet. The CL increases for -1 dB but stays the same for the -3 and -5 dB condition.

System E (predicted F0 and predicted MCC), has a median of 2.5 in terms of naturalness when listening in quiet. It is observed that as the noise levels increased, the listeners reported higher naturalness compared to quiet. Like all other conditions, CL increased as the noise levels increased. Similar to System D, in the -3 and -5 dB conditions the CL remained the same.

System F (full TTS), has a median of 3 in naturalness in most conditions. However, in the -5 dB condition, naturalness was the highest with a median of 4. This was the same observation for System D and E. The CL was reported as medium in quiet and increased to difficult at -1 dB and stayed difficult for all remaining noise conditions.

In all systems except System D, motivations levels were high across all listening conditions except in the -3 dB condition where motivation was the highest. System D sustained motivation equally in all conditions.

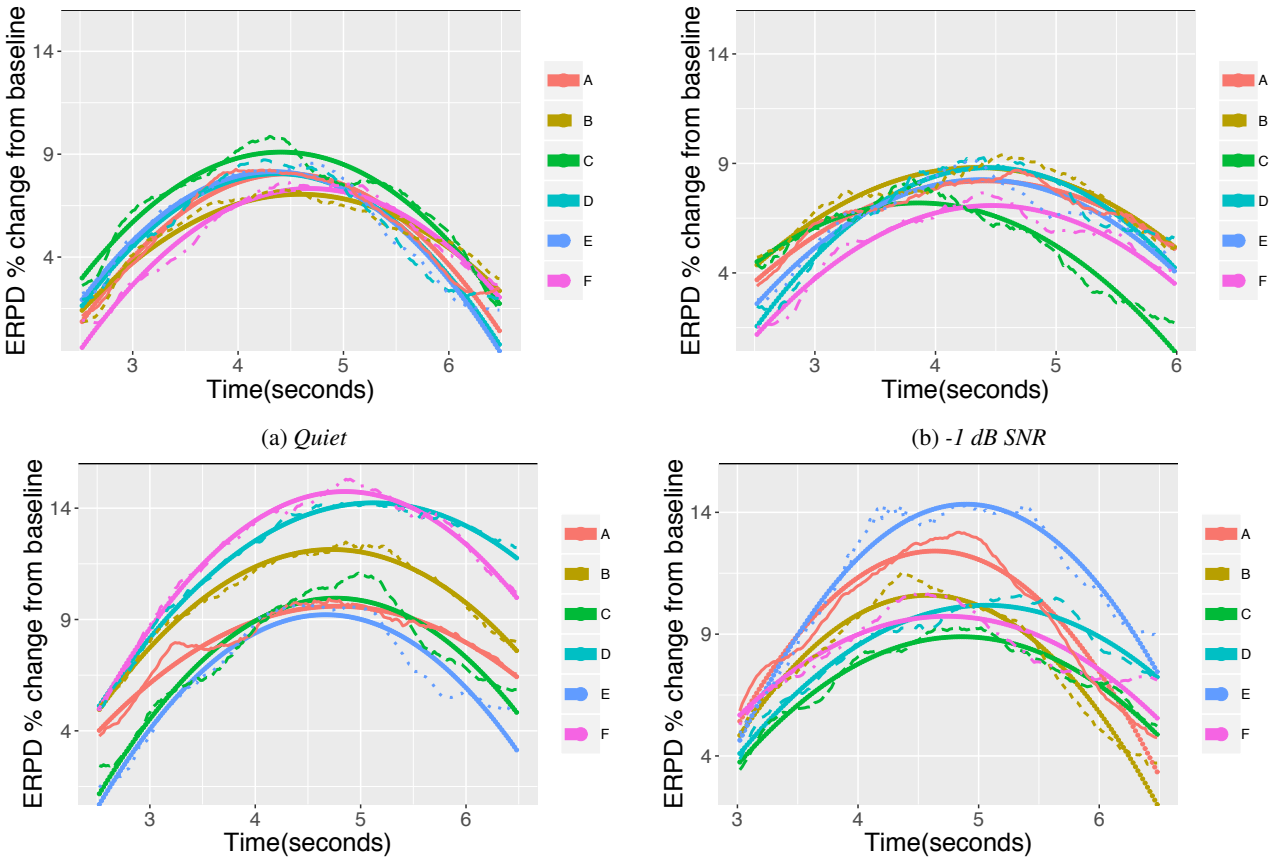
4. Discussion

4.1. Discussion of time terms

The intercept term represents mean pupil dilation. Across all experiments, we observed that the lowest and greatest mean pupil dilation is closely related to the height of the peak pupil dilation for each system. In quiet, System F (full TTS) has the lowest mean pupil dilation (see Figure 1).

If pupil dilation was indexing listening effort in this condition, this would imply that synthetic speech is easier to listen to than natural speech. The self-reported cognitive load scores across all listening conditions show natural speech is perceived as easier to listen to compared to synthetic speech. Therefore, this supports our belief that synthetic speech is not easier to listen to than human speech. Furthermore, in [12], the same result (ie., high quality synthetic speech evokes a smaller pupil dilation compared to natural speech) was observed. It was confirmed that engaged attention is being measured in this case and not listening effort. Since the intelligibility even at -1 dB SNR was still at ceiling and System A (human speech) evoked the highest mean pupil dilation it is likely that the engaged attention was still being indexed by the pupil response at the -1dB condition.

In the -3 dB condition, synthetic speech evokes the greatest



(a) *Quiet*

(b) *-1 dB SNR*

(c) *-3 dB SNR*

(d) *-5 dB SNR*

Figure 1: *ERPD % change from the baseline across all participants and conditions for each listening condition. Dotted: raw data, Solid: quadratic model fit*

pupil dilation. This result correlates more with the self-report scores and thus we are more certain that in the -3 dB condition, listening effort is likely the measure being indexed by the pupil dilation. Since the goals of this work is to understand the cognitive load contributions in terms of the listening effort, we focus on the findings observed in the -3 dB and -5 dB conditions in the next section.

The linear term represents the overall slope, which is the change in pupil response from the start to the end of the trial. System A (human speech) has the flattest slope and System F (full TTS) has the steepest slope for all noise conditions except the -5dB condition. This implies that the amount of cognitive resources utilized when listening to human speech gradually increases over time. It is the opposite when listening to synthetic speech where the peak pupil dilation is reached much faster. However, in the most difficult SNR this is reversed.

The quadratic term represents the *shape* of the peak pupil dilation. When the estimate in the model is close to zero it has a broader shape peak and when the estimate is high in value the peak is sharper. System A has the sharpest peak in quiet and -5 dB which could imply high engagement and high load. For the middle SNR listening conditions it has the broadest peak. System F, however, has the broadest peak shape in quiet, -1 dB and -5 dB and we know that engagement was low and listening effort was high which is opposite to System A. When comparing the shape as we move from one condition to another, we find interesting trends that reveal a relationship to the listening effort when listening in noise.

4.2. Key findings

4.3. Self-reported measures

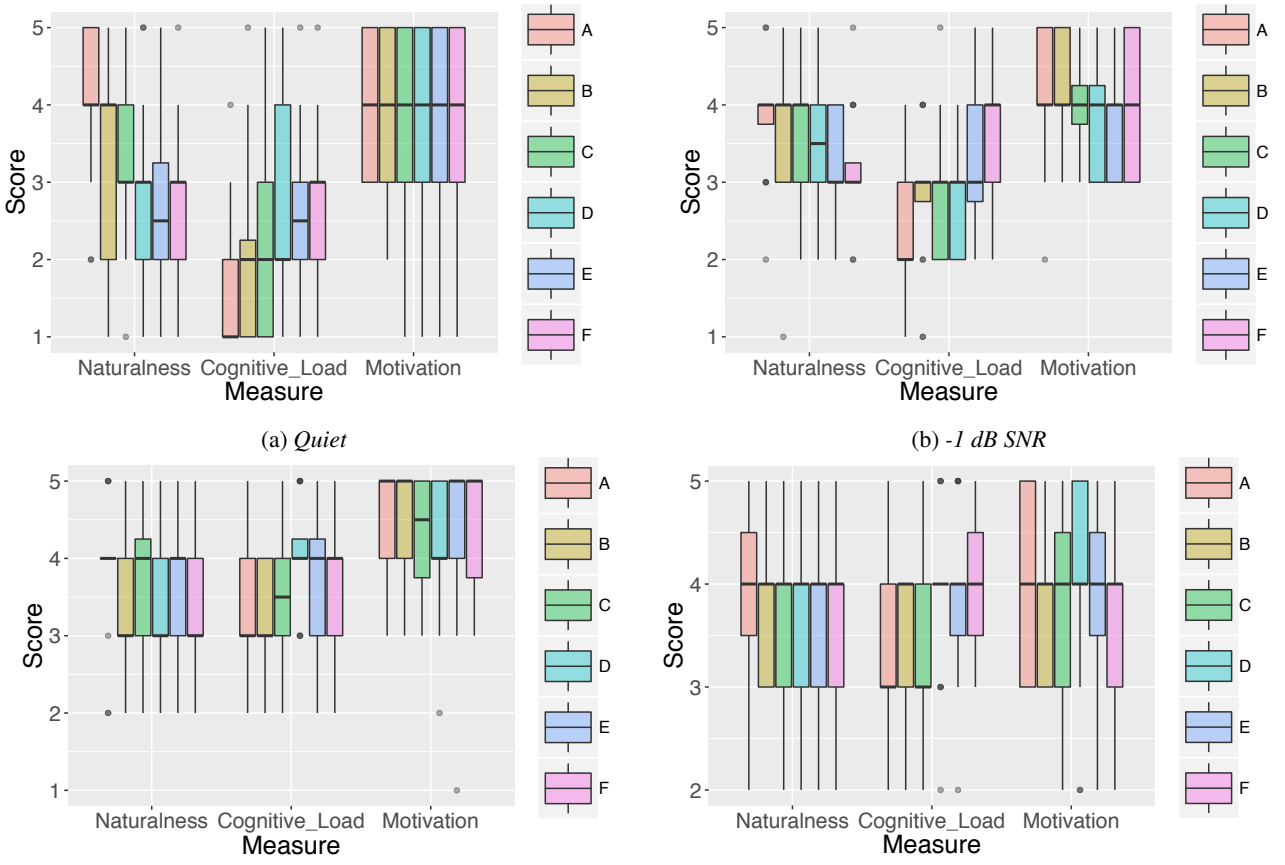
From the self-reported CL we observe that the maximum cognitive load across all listening conditions is reached at the -3 dB condition for Systems A, C, D, and E. Systems A and C have a maximum median of 3 whilst all other systems have a maximum of 4. System F (full TTS) is the only system where the maximum CL is reached at the -1dB condition. System B (vocoded speech) reaches a maximum in the -5 dB condition.

Reported naturalness remained high for Systems A, B and C across all listening conditions. For System D and F, naturalness was only perceived as high in the -5 dB condition. For System E, naturalness was only perceived as high from the -3 dB condition.

4.4. Noisy -3 dB condition

In the -3 dB condition, the ERPDs (see Figure 1) show System A, C and E have a low ERPD. System B falls in the middle and Systems D and F have a high ERPD.

With respect to mean pupil dilation for -3 dB (see Table 3a), System D and F had the highest means. These systems also had the steepest slopes (see Table 3b). System C had the lowest



(a) *Quiet*

(b) *-1 dB SNR*

(c) *-3 dB SNR*

(d) *-5 dB SNR*

Figure 2: *Self-reported measures (1 - very unnatural, very difficult and unmotivated; 5 - very natural, very easy and highly motivated)*

mean and flattest slope. System A and B fell in the middle with respect to their means but were both flat in slope. System E fell in the middle in both cases.

With respect to peak shape for -3 dB (see Table 3c), the peak shape for System A and E remain the same as in the -1 dB condition. All other systems change peak shape. Systems B, C and F change from broad to sharp. In contrast, System D changes from sharp to broad.

Based on the observations described above, a high mean pupil dilation and high ERPD together with a change in peak shape corresponds to a high listening effort. This is observed for Systems D and F. These systems also have the steepest slopes compared to all other systems. This is additionally supported by the self-reported CL scores where a CL of 4 was reached faster than all other systems. Both these systems involve *predicted* spectral features generated by an acoustic model. The difference between them is, System D uses natural duration whilst System F uses predicted duration. This leads us to believe that poor spectral prediction in the acoustic model contributes to an increased cognitive load.

System B evokes a mean pupil dilation that is still on the high end of the scale and changes in terms of its peak shape. Therefore, it is possible that the vocoding itself contributes to an increased cognitive load. The load, however, appears to get compounded by poor spectral prediction which is supported by the mean pupil dilation and ERPD being greater for System D and F than System B. Furthermore, System B has a flatter slope than Systems D and F which is also supported by the self-reported CL reaching a maximum only in the most difficult condition (-5 dB).

In contrast, Systems A and E evoke low ERPD and have mean pupil dilations that falls in the middle. In terms of their peak shapes System A and E remain the same as in the -1dB condition. Low ERPD and an unchanged peak shape is therefore associated with low listening effort. Comparing System E (low listening effort) to System F (high listening effort), they differ only in duration. Therefore, duration prediction contributes to an increased listening effort for synthetic speech. System E, however, received higher CL scores compared to System A. It was also perceived as unnatural in quiet. This indicates that human speech still evokes lower listening effort than synthetic speech even when using *perfect* duration.

System C has a combination of properties that are associated with both high and low listening effort. It evokes a low ERPD, a low mean pupil dilation and the flattest slope. However, we observe changes in peak shape between the -1 dB and -3 dB conditions. It is possible that predicted F0 helps in lowering the load but the conflict with with vocoding MCC causes overall load to still remain high. The perceived naturalness was lower than System A and B in quiet but the CL, like System A only reached a maximum of 3. Therefore predicted F0 may help reduce cognitive load but still sounds unnatural.

4.5. Noisy -5 dB condition

We observe that the mean pupil dilation is the greatest for System A and E. This is in contrast to the previous condition where both these systems evoke the lowest ERPD. Once again, System B evokes a pupil dilation that falls in the middle. All three systems (A, E and B) are also the only three systems to have sharp peaks in this condition whilst all other systems have broad peaks. The high pupil dilation and sharp peak correspond to high listening effort in this condition which is expected given that it is the most difficult noise condition in this work. More interestingly, we observe that Systems D and F evoke a lower mean pupil dilation in this condition compared to the easier noise condition (-3dB). A similar finding was observed in [1] when measuring cognitive load of TTS at -5dB. It was said to indicate a task that is too challenging for the listener. Although Systems A, B and E evoked the greatest responses, in relation to the remaining systems this indicates that the cognitive load was at least still manageable. The only property that Systems A, B and E share is natural duration. However, one can argue that System D also use natural duration yet suffers in this condition. This reveals that it is not duration in isolation that contributes to increased cognitive load but also the correlations that exist between the spectral and F0 parameters which are absent in Systems C and D as the features were extracted/predicted separately. Although Systems C and D are unrealistic, it highlights the importance of modelling spectral and F0 features in a unified framework such that their correlations are kept intact. The ERPD for System C was the only system that remained unchanged and had the lowest mean pupil dilation in this listening condition. It is interesting that System C has the same perceived CL score yet the ERPD and mean remain low. Therefore, this suggests the importance predicted F0 plays in the generation of synthetic speech.

5. Conclusion

The cognitive load of synthetic speech indexed by the evoked pupil response in both quiet and noisy conditions were investigated. Our results confirm that even high quality output generated by a DNN-speech synthesis system still evokes greater cognitive load than natural speech when listening under noisy conditions. Attention and engagement is low when listening in quiet and high cognitive load is reached much faster when listening to synthetic speech than natural speech. The contributions of cognitive load in DNN-based speech synthesis are mainly due to poor spectral prediction and poor duration prediction. When combining speech parameters extracted from natural speech with predicted acoustic features, correlations between these features are destroyed. This alone appears to result in an increased cognitive load. However, this result highlights the importance of modelling spectral, F0 features and duration in a unified framework. Conventional DNN-based speech synthesis models like the one used in this work, models duration and acoustic features sequentially. This could explain why they still evoke high cognitive load compared to natural speech. Sequence-to-sequence models addresses this limitation. Results of System C suggest that predicted F0 plays a role in reducing cognitive load but further investigations need to be done to confirm this result. In conclusion, improved prediction of the spectrum, F0 and of duration will lead to reduced cognitive load in synthetic speech. However, this is for the case of the rather neutral prosody in the data we used and further work is needed to investigate the case of expressive prosody.

6. Acknowledgements

This project has received funding from the EU's H2020 research and innovation programme under the MSCA GA 675324 (the ENRICH network: www.enrich-etn.eu). Thanks to Nina Diviza and Jane Crumlish for their assistance in running the perceptual tests.

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

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

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

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

[6] B. Laeng, S. Sirois, and G. Gredebäck, "Pupillometry: A window to the preconscious?" *Perspectives on psychological science*, vol. 7, no. 1, pp. 18–27, 2012.

[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] 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.

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

[10] 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.

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

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

[13] 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*, vol. 99, no. 7, pp. 1877–1884, 2016.

[14] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system." in *SSW*, 2016, pp. 202–207.

[15] 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.

[16] 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.

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

[18] S. E. Kuchinsky, J. B. Ahlstrom, K. I. Vaden Jr, S. L. Cute, L. E. Humes, J. R. Dubno, and M. A. Eckert, "Pupil size varies with word listening and response selection difficulty in older adults with hearing loss," *Psychophysiology*, vol. 50, no. 1, pp. 23–34, 2013.