THE UNIVERSITY *of* EDINBURGH

THE UNIVERSITY
of EDINBURGH

Evaluating cognitive load of text-to-speech synthesis

Avashna Govender

supervised by
Prof. Simon King and Dr. Cassia Valentini-Botinhao

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2023

Lay Summary

Text-to-speech synthesis is the automatic process of converting text into its speech counterpart. This technology has been increasingly integrated into real-world applications that exists today such as audio-books, language learning applications and personal assistants to name a few. With increasing usage of such applications it becomes important to evaluate the users' experience. One key aspect to consider when evaluating the users' experience is to understand any potential negative implications that may occur when listening to artificially produced speech - if any. In the past, synthetic speech was perceived to be more difficult to process than human speech. In other words, the cognitive load (defined as the deliberate allocation of mental resources) when listening to synthetic speech is much greater in comparison to listening to human speech. Cognitive load, however, was only ever evaluated on rule-based systems that are rarely in use today. The quality in terms of naturalness and intelligibility of synthetic speech has drastically improved since then. Therefore, there was a need to understand how synthetic speech produced by more recent text-to-speech systems interact with our cognitive processing system. This thesis makes two main contributions to the field of evaluating text-to-speech synthesis. In Part I, two approaches were investigated as potential measures for cognitive load: the dual task paradigm and pupillometry. The contribution lies in the proposed methodology for measuring the cognitive load of synthetic speech. To our knowledge, the work presented is the first attempt at using pupillometry to measure the cognitive load of synthetic speech. In Part II, the contribution lies in applying the methodology developed in Part I to measure the cognitive load of current state-of-the-art text-to-speech systems. The aim was to evaluate whether high quality synthetic speech produced today is still perceived to be more difficult to process than human speech. In addition, we take it a step further by setting up the proposed methodology in a manner that enables us to better understand where increased cognitive load contributions come from. The knowledge gained from investigating the cognitive load is crucial for improving the overall users experience as it helps us to make better informed decisions on which aspects of the system need to be improved so that synthetic speech can be optimised for low cognitive load in future.

Abstract

This thesis addresses the vital topic of evaluating synthetic speech and its impact on the end-user, taking into consideration potential negative implications on cognitive load. While conventional methods like transcription tests and Mean Opinion Scores (MOS) tests offer a valuable overall understanding of system performance, they fail to provide deeper insights into the reasons behind the performance. As text-to-speech (TTS) systems are increasingly used in real-world applications, it becomes crucial to explore whether synthetic speech imposes a greater cognitive load on listeners compared to human speech, as excessive cognitive effort could lead to fatigue over time. The study focuses on assessing the cognitive load of synthetic speech by presenting two methodologies: the dual-task paradigm and pupillometry. The dual-task paradigm initially seemed promising but was eventually deemed unreliable and unsuitable due to uncertainties in experimental setups which requires further investigation. However, pupillometry emerged as a viable approach, demonstrating its efficacy in detecting differences in cognitive load among various speech synthesizers. Notably, the research confirmed that accurate measurement of listening difficulty requires imposing sufficient cognitive load on listeners. To achieve this, the most viable experimental setup involved measuring the pupil response while listening to speech in the presence of noise. Through these experiments, intriguing contrasts between human and synthetic speech were revealed. Human speech consistently demanded the least cognitive load. On the other hand, state-of-the-art TTS systems showed promising results, indicating a significant improvement in their cognitive load performance compared to rule-based synthesizers of the past. Pupillometry offers a deeper understanding of the contributing factors to increased cognitive load in synthetic speech processing. Particularly, an experiment highlighted that the separate modeling of spectral feature prediction and duration in TTS systems led to heightened cognitive load. However, encouragingly, many modern end-to-end TTS systems have addressed these issues by predicting acoustic features within a unified framework, and thus effectively reducing the overall cognitive load imposed by synthetic speech. As the gap between human and synthetic speech diminishes with advancements in TTS technology, continuous evaluation using pupillometry remains essential for optimizing TTS systems for low cognitive load. Although pupillometry demands advanced analysis techniques and is time-consuming, the meaningful insights it provides into the cognitive load of synthetic speech contribute to an enhanced user experience and better TTS system development.Overall, this work successfully establishes pupillometry as a viable and effective method for measuring cognitive load of synthetic speech, propelling synthetic speech evaluation beyond traditional metrics. By gaining a deeper understanding of synthetic speech's interaction with the human cognitive processing system, researchers and developers can work towards creating TTS systems that offer improved user experiences with reduced cognitive load, ultimately enhancing the overall usability and acceptance of such technologies.

Note: There was a 2-year break in the work reported in this thesis where an initial pilot was performed in early 2020 and was then suspended due to the covid-19 pandemic. Experiments were therefore rerun in 2022/23 with the most recent state-of-the-art models so that we could determine whether the increased cognitive load result is still applicable. This thesis was thus concluded by

answering whether such cognitive load methods developed in this thesis are still useful, practical and/or relevant for current state-of-the-art text-to-speech systems.

Acknowledgements

I would like to thank the following people:

- My supervisors, Prof. Simon King and Dr. Cassia Valentini-Botinhao, for your advice, guidance and support throughout this journey.
- Marie-Curie Actions for funding my work
- All ESRs and seniors from the ENRICH project for your helpful feedback, knowledge exchange and collaborations especially Anita Wagner for your guidance as my external supervisor.
- Jane Crumlish, Nina Diviza and Ella Crocker for assisting me in the running my listening tests.
- All fellow CSTR researchers for providing such a friendly working environment, for your constructive criticism and suggestions during my internal reviews and conference practice presentations.
- My family in South Africa - Mum, Dad, Shivani and Devashen for your continuous support and motivation to ensure I saw this through. For believing in me on the days I couldn't believe in myself.
- My dearest friends Carol, Lucia, Kate, Janie, Natascia, Olina, Sneha, and Elif for always being there to lend a listening ear and for your support and encouragement throughout this journey.
- Julie Anne for always going above and beyond to help me recruit participants for my listening tests.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Avashna Govender)

Contents

List of Abbreviations

ANOVA Analysis of Variance

ASR Automatic Speech Recognition

CL Cognitive Load

DBN Deep Belief Network

DC-TTS Deep Convolutional TTS

DNN Deep Neural Networks

DSP Digital Signal Processing

EEG Electroencephalogram

ERPD Event-related pupil dilation

F0 Fundamental frequency

FC Fully Connected

FF Feed-Forward

FNIR Functional Near-Infrared Spectroscopy

FMRI Functional Magnetic Resonance Imaging

GCA Growth Curve Analysis

GRU Gated Recurrent Unit

HMM Hidden Markov Models

IQR Inter-Quartile Range

LC-NE Locus Coeruleus - Norepinephrine

LE Listening Effort

LQ-HMM Low-Quality HMM

LSTM Long-Short-Term-Memory

MCC Mel-Cepstral Coefficients

MLSA Mel-Log Spectrum Approximation

MOS Mean-opinion-score

pDCT Proportional Dual Cost Time

PDF Probability Density Function

POS Part-Of-Speech

RBM Restricted Boltzmann Machine

RNN Recurrent Neural Network

RP Received Pronunciation

RT Reaction Time

SNR Signal-to-noise ratio

SMS Semantically Meaningful Sentences

SPSS Statistical Parametric Speech Synthesis

SSRN Super-Spectrogram Resolution Network

STFT Short Time Fourier Transform

SUS Semantically Unpredictable Sentences

TEPR Task-invoked Pupillary Response

TTS Text-to-speech

US Unit Selection

WER Word-error-rate

List of Figures

List of Tables

Chapter 1

Introduction

Speech technologies such as Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) synthesis form the core of human computer interaction. The role of an ASR system is to recognize input speech and convert it to a written text output and TTS does the reverse. Integration of these technologies into real-world applications that exist today such as audio-books, learning language applications, personal assistants and assistive communication aids have transformed the way humans communicate with computers as well as with each other. Thus, such technologies have become part of our everyday lives.

From the standpoint of engineers developing such technologies, it is therefore of utmost importance to ensure that all potential negative implications are accounted for and minimized. Yet efforts in evaluating the existence of such implications are few and far between.

Evaluation methods have remained the same for over a decade. TTS is typically evaluated by computing intelligibility which involves asking the listener to transcribe sentences, or by collecting listeners' opinions on naturalness using Likert rating scales. Such methods are often limited to only these two properties: intelligibility and naturalness. Whilst both these properties remain important for gauging how closely synthetic speech resembles human speech, there are also other properties like users' experience that are equally important yet under-evaluated.

Especially now that the usage of speech technology has become popular, evaluating users' experience is crucial. An important aspect of evaluating users' experience is by understanding the difficulty experienced by the listener - if any - when listening to synthetic speech.

In understanding 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* (defined as the deliberate allocation of mental resources to a given task [Pichora-Fuller et al., 2016]) is required.

Evaluation of cognitive load (CL) of text-to-speech synthesis was last considered when rule-based speech synthesis systems were popular. In those works, it was reported that it is more demanding to process synthetic speech than it is to process human speech [Nusbaum and Pisoni, 1985, Delogu et al., 1998, Winters and Pisoni, 2004]. However, the advancements made in improving TTS systems has drastically improved the level of quality that is produced. So, on one hand, the findings presented in previous works are likely to be outdated but, on the other hand, if such claims hold

true, then as a consequence this puts users at a risk of facing negative implications like fatigue which could ultimately lead to an unpleasant user experience.

In this work, we aim to formally evaluate the cognitive load imposed on listeners when listening to speech produced by TTS.

1.1 Organization of this thesis

This thesis has two main aims and therefore has been divided into two parts. Part I focuses on the measurement of cognitive load when listening to synthetic speech, with the aim of finding a suitable paradigm. The purpose of the chosen paradigm is to detect differences between various TTS systems with respect to how effortful they are to listen to as well as to understand whether synthetic speech demands greater cognitive effort than human speech. Using this knowledge, Part II applies the paradigm developed in Part I to state-of-the-art TTS systems. In addition, we delve deeper with the aim of uncovering the contributions that lead to an increased cognitive load. These contributions are investigated intrinsically by looking at the properties of the the synthetic speech signal itself.

Before addressing each aim, **Chapter 2** provides the relevant background necessary to understand the work presented in this thesis. In Chapter 2, we present the various text-to-speech models that have been evaluated in different parts of this thesis, paying particular attention to the limitations of each model. We then describe existing methods for the evaluation of text-to-speech synthesis and discuss the major shortfall with respect to evaluating the user experience. User experience can be tackled from several angles, but within the scope of this thesis, we are mainly interested in the *difficulty* of listening to synthetic speech. To evaluate difficulty of listening, it is important to understand how synthetic speech interacts with the human cognitive processing system during a listening task. Thus, in this chapter, we discuss existing methods of measuring cognitive load with particular focus on measures of listening effort (LE). Further theoretical background on cognitive processing is provided to understand the fundamental processes that take place during a listening task such that meaningful conclusions can be drawn from the results presented throughout this thesis.

From the survey of cognitive load measurement/methods described in Chapter 2, two methodologies were chosen: the dual-task paradigm and pupillometry. In Part I, each of these methods is discussed, implemented and tested in **Chapter 3 and Chapter 4** respectively. **Chapter 5** compares and summarises the findings of both methods with the aim of motivating the choice of method for measurement of cognitive load that will be applied in the investigations that follow in Part II.

Investigations in Part II were performed under the hypothesis that speech produced by a TTS system is more difficult to listen to than human speech. The approach undertaken was thus to design experiments in a manner that would attempt to test this hypothesis. In **Chapter 6**, we start our investigations by first exploring the properties of the signal itself. At the time of conducting these experiments, statistical parametric speech synthesis (SPSS) was the dominant synthesis paradigm within the TTS research community. In a typical SPSS system, an acoustic model is trained to

generate some speech parameters from a carefully constructed representation of text input which is passed on to a vocoder that is responsible for reconstructing the speech. The representation contained in those speech parameters produced as output from the acoustic model is therefore dependent on the type of vocoder used. As a result, the choice of these acoustic speech parameters plays an important role in the quality of the generated output. Therefore, the experiment was set up to investigate the influence of each vocoder speech parameter in an SPSS system. **Chapter 7** measures the cognitive load of state-of-the-art TTS systems. At the time that the first experiment was conducted, sequence-to-sequence models were dominating the TTS field and therefore these models were evaluated. Since there was a 2-year break in the thesis, newer models were developed and therefore it is important for us to understand whether speech produced by the latest text-to-speech models still demand more cognitive resources than human speech. **Chapter 8** summarises and discusses all findings in Part II. Finally, **Chapter 9** concludes the thesis with direction of future work.

1.2 Research Questions and Hypotheses

Based on the above mentioned aims, the following research questions will be addressed in this work:

1. How can we measure the cognitive load when listening to synthetic speech and can a suitable paradigm be developed that is capable of detecting differences in the cognitive effort required to listen to various TTS systems?
2. Does synthetic speech demand greater cognitive effort compared to human speech?
3. If cognitive effort is greater than human speech, what are the contributing factors that lead to increased cognitive load in synthetic speech processing?
4. With the recent advancements made in TTS, do modern TTS systems still demand greater cognitive load than human speech?
5. Is the research conducted in this thesis still relevant?

The hypotheses for each of the research questions are as follows:

1. There are various methods proposed in the literature that measure the cognitive load of human speech and rule-based speech synthesizers, some which include behavioural methods like the dual-task paradigm and others that include physiological methods such as skin conductance and pupillometry to name a few. In this work, the dual-task paradigm and pupillometry paradigm were investigated as two potential methods for measuring the cognitive load of synthetic speech. The hypothesis is that pupillometry will prove to be the more viable method as it has consistently shown promising results in several recent works regarding the cognitive load of human speech. However, it will be worthwhile to investigate the dual-task paradigm

first as it is a simpler method to implement than pupillometry. Therefore, either method could potentially be a suitable but what will set them apart is their sensitivity to detecting cognitive effort differences between various systems compared. Since pupillometry is a method that the listener is not directly in contol of compared to behavioural methods, the hypothesis is that pupillometry will be the measurement that is more sensitive in detecting cognitive load difference between the two methodologies investigated.

2. The hypothesis is that older TTS systems such as Unit Selection and Hidden Markov Model (HMM)-based models will demand a greater cognitive load than human speech but what we anticipate observing is that over time as TTS systems improve, the cognitive load will reduce and the effort demanded by more recent TTS systems will eventually converge with human speech.

3. We hypothesis that the contributing factors to increased cognitive load would be due to poor acoustic feature predictions in older traditional models such as SPSS models. However, we expect that modern TTS systems which are considered to be indistinguishable from human speech (in naturalness and intelligibility) will not differ from human speech.

4. The hypothesis is that state-of-the-art TTS systems that exist today will demand a similar amount of cognitive load compared to human speech and possibily demand less cognitive load than human speech.

5. The work conducted in this thesis still remains relevant as evaluating the cognitive load of TTS systems is still necessary to understand the users experience when using real-world applications with TTS embedded in them.

1.3 Published work

At the start of this work, I set milestones in the form of publication targets that I continually strived towards. As a result, as part of this work I have written a number of peer-reviewed papers that have been published. Each of these papers has been rewritten as parts of chapters in this thesis. [1]

- Govender, A., and King, S (2018). Measuring the cognitive load of synthetic speech using a dual-task paradigm. In Proc. Interspeech 2018, Hyderabad, India.

- Govender, A., and King, S (2018). Using pupillometry to measure the cognitive load of synthetic speech. In Proc. Interspeech 2018, Hyderabad, India.

- Govender, A., Wagner, AE., and King, S (2019).Using pupil dilation to measure cognitive load when listening to text-to-speech in quiet and in noise. In Proc. Interspeech 2019, Graz, Austria.

[1] As time passes we grow and we learn and therefore my initial ideas, thoughts and beliefs described in these papers, have too evolved.

- Govender, A., Valentini-Botinhao, C., and King, S (2019). Measuring the contribution to cognitive load of each predicted vocoder speech parameter in DNN-based speech synthesis. In Proc. 10th Speech Synthesis Workshop 2019, Vienna, Austria.
- Govender, A., and King, S (2023). Cognitive load of modern TTS systems under noisy conditions, Cognitive AI workshop 2023, Bari, Italy)

Chapter 2

Background

This chapter provides the theoretical background related to the three central topics of this thesis: 1) text-to-speech synthesis, 2) evaluation and 3) cognitive load. Section 2.1 introduces the topic of text-to-speech synthesis followed by an overview (in chronological order) of the various TTS systems evaluated in this thesis in Section 2.2. Next, the perception of synthetic speech is discussed in Section 2.3 followed by a discussion on current evaluation methods used for TTS and the limitations of such methods in Section 2.4. Finally, Section 2.5 defines cognitive load in relation to listening effort and a discussion of common cognitive load measurement methods.

2.1 Text-to-speech synthesis

Text-to-speech synthesis artificially produces human speech by automatically converting the written text form into its spoken counterpart in a desired accent, language and voice. This conversion process in older and more traditional TTS models typically has two main components: a front-end and back-end. The front-end is responsible for the processing of the text input. This input is transformed into a meaningful representation that is passed to the back-end that is responsible for the generation of the corresponding speech waveform.

The most simplistic structure of the front-end, shown in Figure 2.1 typically comprises modules including text normalisation, part-of-speech (POS) tagging and letter-to-sound rules whilst more advanced front-ends also include modules designated to predicting prosody. Prosody is the term used to describe the rhythm, stress or intonation of speech. In other words, the tone required to deliver the speech output appropriately. In the text normalisation module, the input text is converted such that numerical values are expanded into their text counterpart and abbreviations are expanded into full words. The words in the text are then passed to a POS tagger which helps the front-end disambiguate pronunciations. Most well-resourced languages have lexicons that define words in the language as a sequence of phonemes. A phoneme is the smallest unit of sound in a language that describes categories of speech sounds that have the same linguistic function and represent the various ways a specific sound can be pronounced. Therefore, the lexicon is an important resource that aids the front-end in transcribing how to pronounce words correctly. In the instance that a word does not exist in the lexicon, the letter-to-sound rule module contains carefully hand-crafted rules

that guide the front-end in determining the most appropriate pronunciation. Once an appropriate pronunciation is determined, the phonemes together with their positional and contextual information is captured and referred to as a linguistic specification. This linguistic specification is then passed to the back-end for waveform generation. In the more advanced front-ends, this linguistic specification is used to predict prosody. For example, prosody can be predicted from the text based on specific words that relate to the emotion and specific emotions have specific intonation associated with them. Such features are included in the specification before being passed on to the back-end. It is evident that these modules are heavily language-specific and therefore require extensive linguistic expert knowledge. Fortunately, the text-to-speech synthesis pipeline is constructed such that all natural language processing takes place in the front-end which allows the back-end to operate independently of language-specific knowledge.

Figure 2.1: Illustration of a simple front-end system

The back-end receives the linguistic specification and then generates the corresponding speech waveform using a method like unit selection that is concatenative or by training a statistical generative acoustic model that is coupled with a vocoder. The vocoder is responsible for reconstructing the speech waveform from the parameters generated by the acoustic model. In general, it is the implementation of the back-end which differentiates the various TTS systems. Modern state-of-the-art TTS systems, however, do not always consist of a front-end and back-end separately. Nowadays, end-to-end TTS systems exist in which all modules are collapsed into a single unified framework. Such systems will be described later on in this chapter. A brief overview of each TTS system evaluated in this thesis is provided in the sections that follow.

2.1.1 Concatenative speech synthesis

Concatenative speech synthesis dominated the field for quite some time in the past due to its ability to synthesise close to human-like sounding speech that is highly intelligible [Hunt and Black, 1996, Black and Taylor, 1997, Conkie, 1999, Beutnagel et al., 1999, Schwarz, 2000] Concatenative synthesis is a technique that joins or concatenates pre-recorded and segmented speech units to recreate a speech waveform. It is able to maintain high quality as the speech units typically remain unprocessed. Thus, the output speech contains all characteristics of the human speaker that was originally recorded [Taylor, 2009].

Figure 2.2 illustrates the architecture for a simple unit selection system which is the most common concatenative speech synthesis approach. Input text is provided to the front-end (Figure 2.1) where the text-analysis module produces a linguistic specification. The back-end then uses the linguistic information to look-up recorded speech units in a speech database that corresponds with

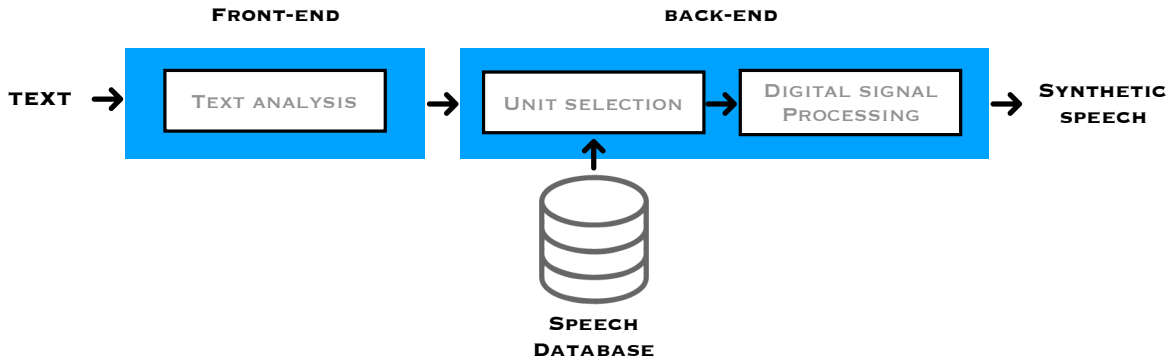


Figure 2.2: Illustration of a simple unit selection system

the linguistic specification. Digital signal processing (DSP) is applied to connect and smooth units to produce the output speech waveform.

Success of producing high quality speech output with this approach is therefore heavily dependent on the speech database available. Typically, the speech recordings are segmented into small units like phones, and di-phones. These units are then labeled with their segmentation and acoustic parameters (fundamental frequency, pitch, duration and energy), positional and contextual information and the next or previous phones. The back-end is responsible for finding the most appropriate sequence of units in the database that matches the information captured in the linguistic specification. However, it is impractical for a database to contain every possible unit in every possible context. Therefore, a major downfall in this approach is revealed at run-time when units of a given utterance to be synthesized do not already exist in the database. In the instance where no match is found, missing units are substituted by similar units found in the database and as a consequence this could lead to deteriorated quality. Efforts in this approach are therefore geared towards being able to optimally select a unit that minimizes the amount of loss in speech quality. This loss in quality can be a result of mismatched targets or bad joins. Therefore, a cost function is used to calculate a target and join cost that is then minimized. The target cost is essentially the selection cost which takes into consideration all possible units that are similar to the target unit being searched for. The join cost takes into consideration the smoothness between the selected units to be connected [Kayte et al., 2015] which relate mostly to the acoustic properties of the unit. Once all units are selected, digital signal processing is performed to concatenate the units to create the output synthetic speech. Sometimes additional processing techniques are applied to smooth the joins between the units. However, the more manipulations done to the unit, the higher the risk of introducing unwanted artefacts to the synthesized speech. Thus, selecting units with low costs means minimal manipulation to the signal which is necessary to make the sentence sound as natural as possible. It is also unlikely that a speech database will capture all possible units necessary for synthesis and thus in most cases the substitution approach is taken. Poor joins are also perceptually more noticeable by listeners which leads to sub-optimal quality.

A common approach to minimize the cost function is by using probabilistic methods. Such unit selection systems that extend to using probabilistic models are called hybrid synthesis. Hybrid

synthesis is therefore a combination of unit selection and SPSS. We will discuss hybrid synthesis later on in this chapter after SPSS has been discussed.

Unit selection requires a lot of memory to store large amounts of speech data and therefore makes it an expensive method. Another drawback is that the output speech will strongly resemble the quality of the recorded database. Therefore, high quality, professional studio recorded speech data is needed. Furthermore, such systems are limited to a single speaker. In the instance that a different language, accent or style of speech is required, a new database will need to be recorded which limits scalability and increases cost.

The unit selection system evaluated in this thesis was one that was submitted to the 2011 Blizzard Challenge [King and Karaiskos, 2011].[1]

2.1.2 Statistical parametric speech synthesis

Statistical parametric speech synthesis is a model-based approach. Unlike unit selection, this type of synthesis generates speech entirely from scratch by learning to model speech parameters [King, 2011, Black et al., 2007, Zen et al., 2009, Yu and Young, 2010]. Since only the models are needed for SPSS systems, they do not require large amounts of memory for storage unlike unit selection. Models utilize less memory than the storage of entire speech recordings. Another advantage SPSS models have over unit selection is that the model can be modified in various ways. Therefore, the model is not limited to sounding only like the speaker in the training data.

To train a statistical model, the speech needs to be parameterised. Therefore, a statistical model cannot function on its own to reconstruct the speech. It works in conjunction with a vocoder that performs analysis, modification and resynthesis. During analysis, features are extracted from the speech that are required for training the model and the resynthesis does the reverse by taking the speech features generated by the model and reconstructing the speech waveform.

HMM-based speech synthesis

At the time when SPSS became popular, Hidden Markov Models (HMMs) were used.

Each observed phone in the data is described by a HMM. Using HMMs, it is assumed that at least one of the models has generated the data, and therefore the model we want is the one with the highest probability. The HMM (used for TTS) models two important components derived from the human speech production system. One part models the vocal tract (spectral features) and the other part models the vocal source (excitation features) [Taylor, 2009]. In Figure 2.3 a conventional HMM-based speech synthesis system is shown. The system consists of two stages: training and synthesis.

Training: In the training part, parameter extraction is first performed to extract spectral and excitation parameters from the speech database. The spectral features, typically mel-cepstral coefficients, are obtained by mel-cepstral analysis [Fukada et al., 1992]. MCCs are features that capture

[1] Under privacy regulations of the Blizzard Challenge, the identity of the selected system has to remain anonymous.

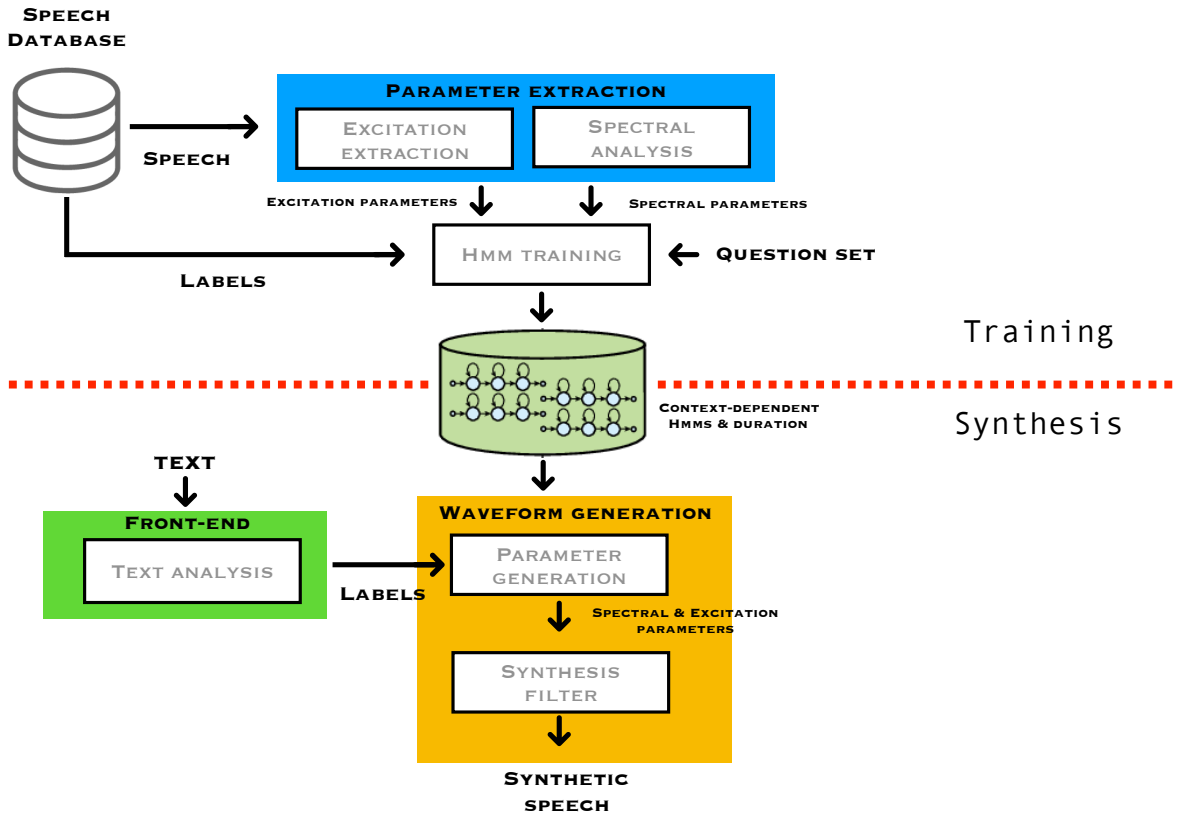


Figure 2.3: Illustration of a conventional HMM-based speech synthesis system

essential characteristics of the spectral content in the speech signal and are important for modeling the phonetic information that is present in the speech signal. The excitation parameters comprise of extracted fundamental frequency. Fundamental frequency is a feature that represents the source of sound and therefore captures the vibrations of the vocal cords which is needed to model natural sounding speech. Since speech evolves as a function of time, dynamic features referred to as the delta and delta deltas, which are the first and second order derivatives, are combined together with the respective static features into single feature vectors that are used to train the HMM model as shown in Figure 2.4. Therefore, the feature vector of the HMMs consist of two streams, one for the spectral part and one for the excitation part. Duration is modelled within the HMM model by splitting the model into three states which represent different parts of a given phone: the beginning, middle and end. A phone's state transition probabilities govern the durational characteristics of the phone. Transition probabilities tell us the probability of moving from one state to the next. Therefore, if the transition probabilities are high for a transition that leads to itself, then it is likely that more observations will be generated by that phone [Taylor, 2009]. Fundamental frequency is modelled using multi-space probability distribution HMMs [Tokuda et al., 1999] and to model the duration of each state multi-dimensional Gaussian distribution is used which captures the statistical properties of the state durations [Tokuda et al., 1998]. All distributions are clustered individually by using a decision tree context clustering technique shown in Figure 2.5. The decision tree context clustering technique is used to cluster the context-dependent HMMs and takes into account phonetic, linguistic, and prosodic contextual factors which affect spectrum, pitch and duration such

as phone identity, stress-related and location factors. These context-dependent HMMs are used to represent distributions of acoustic features given their corresponding linguistic features. Clustering context-dependent HMMs for acoustic modelling is required because the number of contexts covered is constrained to the examples that exist in the training data and this may not be sufficient enough for robust modelling. Therefore, HMMs are grouped by traversing through a binary decision tree and when the terminal node is reached, parameter distributions of those related contexts are shared. Questions asked in the decision tree are those that yield the largest log likelihood gain of the training data. The tied context-dependent HMMs (λ) are then re-estimated using embedded training based on the maximum likelihood (ML) criterion [Tokuda et al., 2000]:

$$\lambda = \arg\max_{\lambda} p(\mathbf{y}|\mathbf{x}, \lambda) \tag{2.1}$$

where $p(.)$ denotes the probability distribution, $y = [y_1^\intercal, y_2^\intercal, y_3^\intercal]$ denotes acoustic features with $\intercal$ frames and y_t is the acoustic feature at frame t, $x = \{x_1, x_2 \ ... \ x_N\}$ is a sequence of linguistic features that correspond to y.

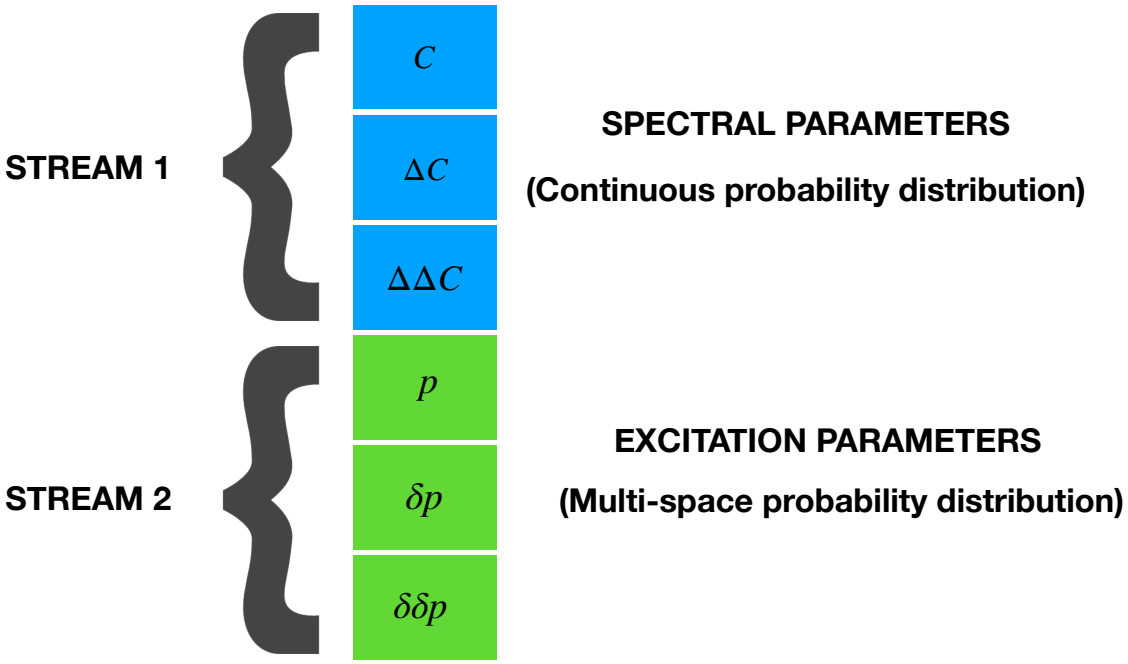


Figure 2.4: Feature vector for HMM training adapted from Yoshimura et al. [1999]

Synthesis: During synthesis, the input text given to the synthesiser is converted to a context-dependent label sequence called the linguistic specification created by the front-end. Then an utterance HMM is constructed by concatenating context-dependent HMMs according to the linguistic specification. Then state durations of the utterance HMM are determined based on the state duration densities. A parameter generation algorithm subsequently generates the sequence of spectral and excitation parameters that maximize their output probabilities [Tokuda et al., 1995]. Finally, a synthesis filter such as the mel-log spectrum approximation (MLSA) filter is used to synthesise the speech waveform [Fukada et al., 1992].

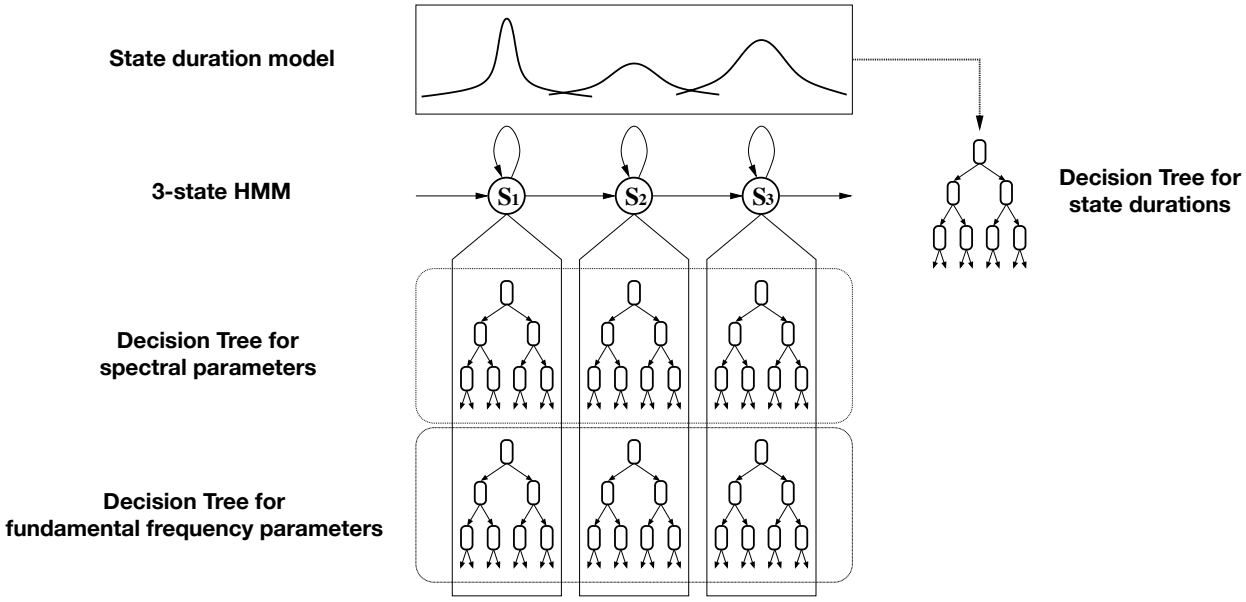


Figure 2.5: Decision tree-based clustering technique adapted from Yoshimura et al. [1999]

Decision tree-clustered context-dependent HMMs work reasonably well but a disadvantage of such an approach is that it divides the input space. Fragmenting training data leads to overfitting and as a consequence the quality of the synthesized speech is degraded [Zen et al., 2013]. Therefore, the acoustic modeling performed using tree-based context clustering, which learns the relationship between linguistic and acoustic features in HMM-based speech synthesis, limits the naturalness of synthetic speech. Furthermore, the clustering technique applied results in averaging all observations in the training data that are associated with a given terminal node in the decision tree. Whilst averaging improves the robustness of acoustic modelling, important speech characteristics are lost in this process by over smoothing of the spectral parameters which leads to a muffled voice quality. Furthermore, decision trees are inefficient in being able to capture complex dependencies that exist between linguistic and acoustic features. Limitations of HMM-based speech synthesis were addressed by re-introduction[2] of neural networks due to increased computational capacity.

Hybrid speech synthesis

Hybrid synthesis systems combine the benefits of SPSS and Unit Selection systems. In unit selection, by concatenating natural speech units, high speech quality is achieved. However, in the instance that speech units do not exist in the database, a cost function is computed to aid the system in selecting the closest matching units for concatenation. In Hybrid speech synthesis, selection of units are informed using statistical models that have been learned from the data. The Hybrid system evaluated in this thesis is a HMM-based unit selection speech synthesis system that uses HMMs trained on acoustic features for phone unit selection. This system[3] was submitted to the 2011 Blizzard Challenge [King and Karaiskos, 2011]. At the time of the 2011 Blizzard Challenge

[2]Neural networks were first introduced in the 1990's

[3]Under privacy regulations of the Blizzard Challenge, the identity of the selected system has to remain anonymous

the use of HMMs dominanted SPSS.

DNN-based speech synthesis

Neural networks were inspired by the human speech production system which is believed to have layered hierarchical structures that transform linguistic information to speech [Yu and Deng, 2010]. In the 1990s, neural networks had already been used to learn the relationship between linguistic and acoustic features but deep learning algorithms required a substantial amount of computational power for training. The computing hardware available in the 1990s was significantly less powerful than the hardware that exists today and thus training deep neural networks efficiently then, was a challenge.

With increased computational capacity in recent times, neural networks have become popular again. Moreso, deeper neural structures can be trained with greater amounts of training data which are referred to as Deep Neural Networks (DNNs). DNNs are complex and far more powerful acoustic models than decision trees and thus have replaced HMMs in SPSS, leading to a rapid improvement in the output speech quality. In early DNN architectures, the DNN replaced only the acoustic modelling aspect of HMM-based speech synthesis which was previously performed using decision-tree clustering techniques. The weights of the DNN are trained using pairs of input and output features extracted from training data. The input features are the linguistic specifications of text in the training data that is produced by the front-end. The linguistic specification includes binary answers to questions about their phonetic, linguistic, and prosodic context. The output features are typically frame-by-frame acoustic features extracted directly from the speech waveforms in the training data. Typically, these features include spectral features like mel-cepstral coefficients, excitation features like fundamental frequency (F0) and energy features like aperiodicity as well as their respective dynamic features which are the first and second derivatives. Since F0 is only present in voiced regions of the speech signal, an additional binary feature is included which denotes whether the current frame is voiced or unvoiced. The training of a DNN acoustic model involves mapping the input features to the output features by representing the conditional probability density function (PDF) of output acoustic features given the input features [Ling et al., 2015]. All acoustic features are trained simultaneously in a unified framework and therefore the complex dependencies between linguistic and acoustic features can be learnt efficiently.

At synthesis time, the conditional distribution of the output acoustic features given the input features of the text to be synthesized can be derived from the trained acoustic models. These output acoustic features are predicted from the conditional distribution under a criterion such as maximizing the output probability. The predicted acoustic features are then sent to a vocoder which is responsible for reconstructing the corresponding speech waveform.

There are several proposed architectures for acoustic modelling using DNNs such as restricted Boltzmann machines (RBMs) [Ling et al., 2013], deep belief networks (DBNs) [Zen and Senior, 2014], feed-forward (FF) neural networks [Zen et al., 2009] and long short-term memory (LSTM) recurrent neural networks (RNNs) [Fan et al., 2014]. In this thesis, FF and LSTM recurrent neural networks were investigated as these are the most common architectures in the literature applied in

text-to-speech synthesis.

A FF neural network is the simplest type of network and is shown in Figure 2.6. Linguistic features are mapped to acoustic speech parameters (vocoder parameters) frame by frame without considering the sequential nature of speech. This DNN-based approach assumes that each frame is sampled independently. Regression is performed via several hidden layers containing units called neurons and each neuron contains an activation function that performs a non-linear computation using the following equation [Wu et al., 2016]:

$$\mathbf{h}_t = H(\mathbf{W}^{xh}\mathbf{x}_t + \mathbf{b}^h) \tag{2.2}$$

where H(·) is a non-linear activation function in a hidden layer, W^{xh} is a weight matrix and b^h is a bias vector. The connection that exists between each of the neurons is its connection weight. Each of the neurons in the previous layer is fully connected to each of the neurons in the current layer and therefore has several connection weights which are captured in the weight matrix. During forward propagation, the weight matrix is multiplied by the activation of the previous layer. A bias is added and the result is propagated to the next neuron. At the final output layer of the DNN is a linear layer that sums all results of all hidden layers performing non-linear transformations from the input:

$$\mathbf{y}_t = \mathbf{W}^{hy}\mathbf{h}_t + \mathbf{b}^y, \tag{2.3}$$

where W^{hy} is the weight matrix, b^y is the bias vector, and $W^{hy}h_t$ is a linear regression to predict target features from the activations in the preceding hidden layer.

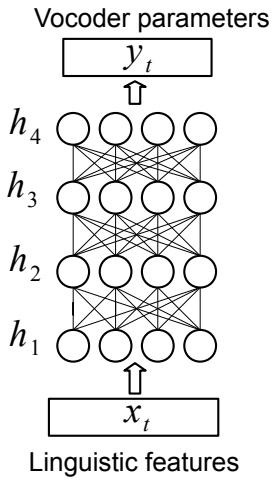


Figure 2.6: An illustration of a simple feed-forward neural network with four hidden layers taken from Wu et al. [2016]

In contrast, RNNs by design are more suitable for sequential data like speech as they have the ability to make predictions by using outputs from previous steps whilst feed-forward neural networks do not. In a RNN, each neuron is not only fully connected to neurons in the previous layer but are

also connected to neuron activations from the previous frame. Therefore, dependencies are formed between the current and previous frames. These RNNs are called uni-directional RNNs. One can also take into account both the previous and next frames in bi-directional RNNs [Fan et al., 2014].

Standard RNNs struggle to model long-term dependencies. To address this problem, the Long-short-term memory (LSTM) network was introduced [Hochreiter and Schmidhuber, 1997]. It contains memory cells with self-connections that are able to store the temporal state of the network. In addition they comprise gates that control the flow of information. Figure 2.7 shows a standard LSTM unit. The input signal and hidden activation of the previous time instance is passed through an input gate, forget gate, memory cell and output gate to produce the activation and is computed using the following equations:

$$\mathbf{f}_t = \sigma(\mathbf{W}^f \cdot [\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}^f), \tag{2.4}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}^i \cdot [\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}^i), \tag{2.5}$$

$$\tilde{C}_t = \tanh(\mathbf{W}^C \cdot [\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}^C), \tag{2.6}$$

$$C_t = \mathbf{f}_t * \mathbf{C}_{t-1} + \mathbf{i}_t * \tilde{C}_t, \tag{2.7}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}^o \cdot [\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}^o), \tag{2.8}$$

$$\mathbf{h}_t = \mathbf{o}_t * \tanh \mathrm{C}_t, \tag{2.9}$$

where i_t, f_t, and o_t are the input, forget, and output gates, respectively; c_t is the memory cell; h_t is the hidden activation at time t; x_t is the input signal; W* are the weight matrices applied on input, b* are the biases.

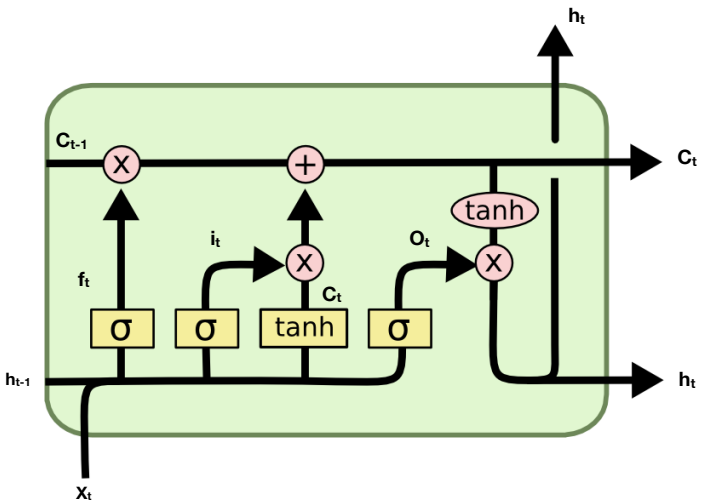


Figure 2.7: An illustration of a LSTM unit taken from Varsamopoulos et al. [2018]

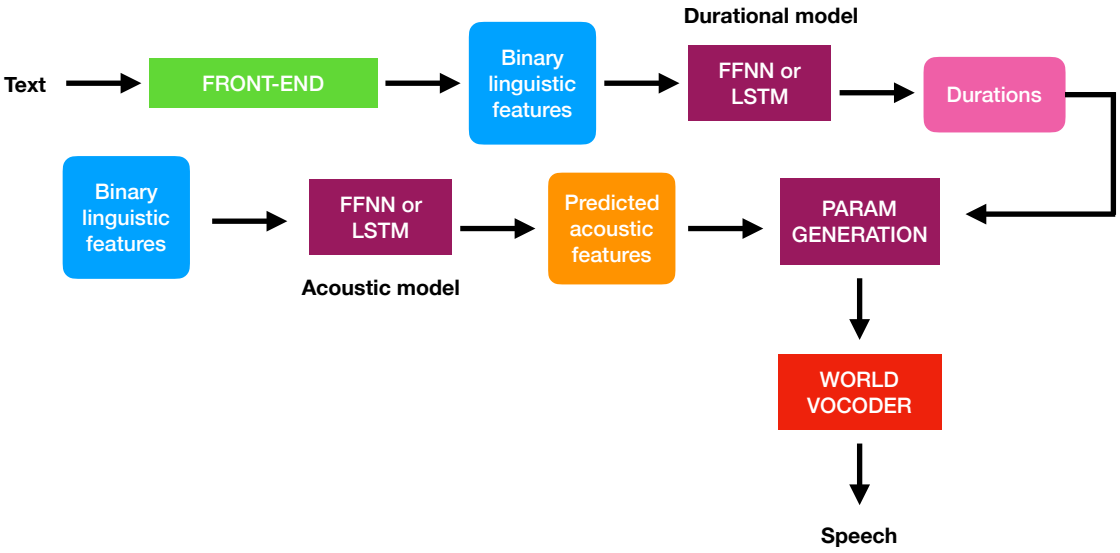


Figure 2.8: Diagram of the Merlin SPSS architecture

The implementation of the FF and LSTM DNNs implemented in this thesis follows the recipes in the Merlin Toolkit [Wu et al., 2016], and all have the architecture shown in Figure 2.8. At synthesis time, text is first converted into a binary linguistic representation which is passed as input to the already trained duration model consisting of either a FF or LSTM DNN. Similarly, the binary linguistic representation is passed as input to the already trained acoustic model also consisting of either a FF or LSTM DNN which produces acoustic features that are used together with the predicted durations to generate the speech parameters needed by the vocoder to reconstruct the speech. The standard vocoder in Merlin is the open-source WORLD vocoder [Morise et al., 2016].

Although traditional DNN-based SPSS architectures like those implemented in [Wu et al., 2016] are capable of producing high quality speech, there are still a few limitations that exist in its implementation. Since such systems have a pipeline architecture, errors may occur which then get propagated through to the output, resulting in deteriorated speech quality. Also, these systems rely on a front-end for natural language processing which typically requires extensive expert domain knowledge. Therefore, using these traditional DNN-based SPSS architectures, high quality synthetic speech is reliant on having access to hand-engineered domain knowledge front-ends. Another limitation is that an alignment between the linguistic features and acoustic features need to be determined ahead of training using a separate forced alignment technique, and duration modelling is also done separately to the acoustic modelling. Poor duration modelling directly affects the output speech quality. Such DNNs also predict adjacent acoustic frames independently and therefore dependencies that exist between frames are lost [Watts et al., 2019].

2.1.3 Sequence-to-sequence-based speech synthesis

Modern sequence-to-sequence neural TTS systems provide close to human speech quality and currently dominate TTS research. These systems do not have the traditional modular approach as described earlier but instead has an architecture where most modelling occurs within a single

unified framework. In such systems, a sequence of linguistic features (characters or phonemes) is converted to a sequence of frame-wise spectral acoustic features (a mel-spectrogram) which are passed to a neural vocoder for synthesis. In this thesis, two architectures of sequence-to-sequence-based speech synthesis were evaluated: Tacotron [Shen et al., 2018] and DC-TTS [Tachibana et al., 2018]. The Tacotron model used in this thesis is a modified version of the original Tacotron model. It is an end-to-end generative model that synthesizes frame-wise mel-spectrograms directly from phonemes. Its architecture comprises pre-nets, an encoder, a decoder coupled with an attention mechanism that is composed mostly of RNNs and a post-processing net. It is referred to as end-to-end as it collapses the traditional text-to-speech synthesis pipeline of having a front-end and back-end into a single neural network. This model is said to produce natural sounding speech that is difficult to distinguish from real human speech. Figure 2.9 shows the architecture for the Tacotron model used in this thesis. Text is first typically converted into characters or phonemes which is fed to the encoder. The encoder is responsible for compressing this linguistic information into a robust sequential representation which is referred to as a character embedding. A content-based attention decoder in Vinyals et al. [2015] is used which learns the mapping between the linguistic information and the acoustic information. A fully-connected output layer is then used to predict the decoder targets. Finally, a post-processing net is tasked with converting the decoder targets to a representation that corresponds with requirements of the neural vocoder that is then used to synthesize targets into speech waveforms. Pre-nets are used to aid with convergence and generalization of the model.

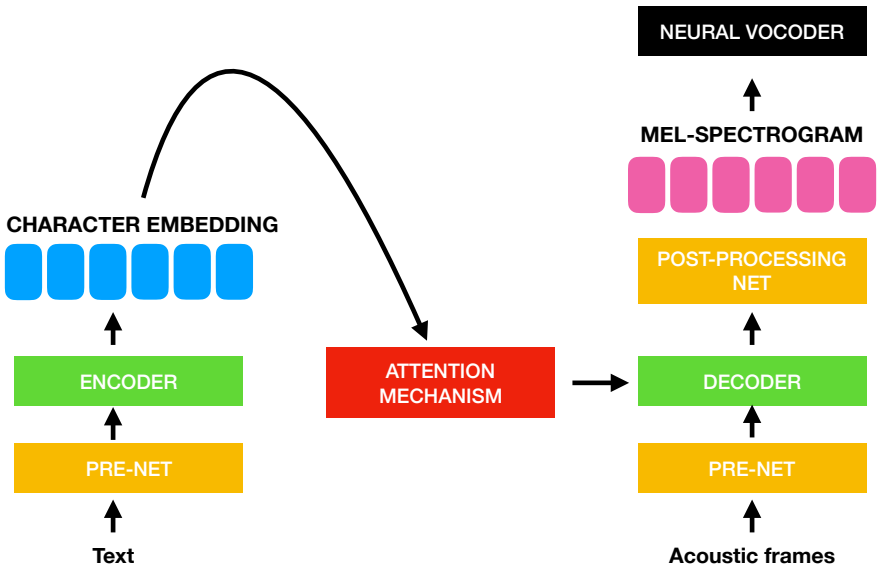


Figure 2.9: Diagram of the typical Tacotron architecture, adapted from Wang et al. [2017]

DC-TTS is also an end-to-end generative model that synthesizes frame-wise mel-spectrograms directly from text. The main differences between DC-TTS and Tacotron is that DC-TTS is fully convolutional as opposed to Tacotron that comprises mostly of recurrent units. Recurrent units are costly to train because they are impractical without expensive machine power and take long periods of time to train. Convolutional-based neural networks can be trained much faster than

RNNs. The DC-TTS model architecture is shown in Figure 2.10. First, text and mel-spectrograms are passed to separate encoders, a text encoder that compresses the linguistic information and audio encoder that compresses the acoustic information. The outputs of the encoders are passed to an attention mechanism where the text and audio is aligned. Guided attention proposed in Tachibana et al. [2018] is applied in DC-TTS as opposed to a content-based attention mechanism used in Tacotron. Then an audio decoder module similar to the decoder in Tacotron generates the target features. Then, a spectrogram super-resolution network (SSRN), similar to the post-processing net of Tacotron, converts the coarse mel-spectrogram generated by the decoder to the full spectrogram which is passed to a neural vocoder for synthesis. Both DC-TTS and Tacotron models used in this thesis are open-source implementations in Watts [2019] andFatchord [2019] respectively.

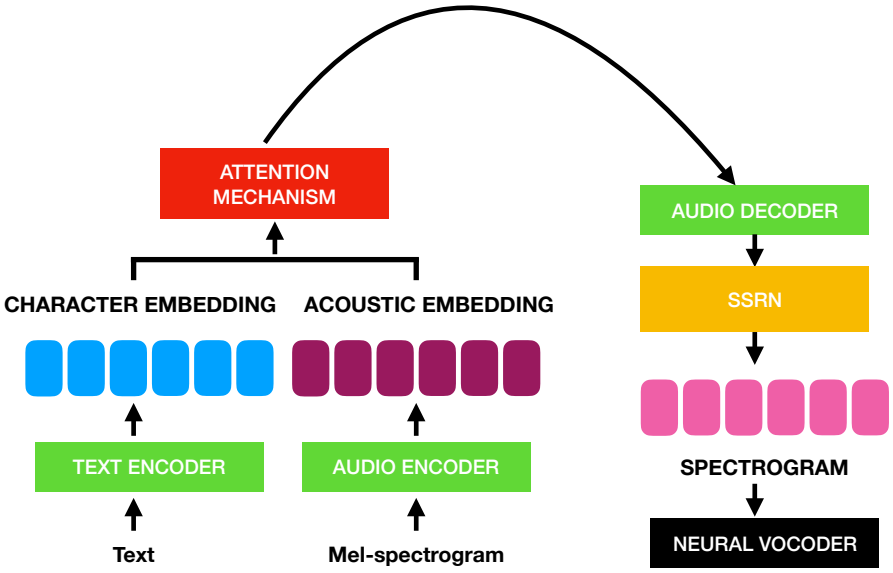


Figure 2.10: Diagram of the DC-TTS architecture, adapted from Tachibana et al. [2018]

Existing parametric models generate speech by passing their outputs through vocoders for reconstruction. Vocoders in the past were heavily signal processing-based such as STRAIGHT [Kawahara, 2006] which was designed to model the source and filter model derived from the human speech production system. A neural paradigm was introduced called Wavenet which is capable of directly modelling a raw waveform through sampling that is able to produce high fidelity speech output [Van Den Oord et al., 2016]. This lead to an uprise in research involving neural vocoding such as Parallel Wavenet [Van Den Oord et al., 2016], WaveGlow [Prenger et al., 2019] and WaveRNN [Kalchbrenner et al., 2018] which are all optimized variations of Wavenet. Wavenet is a convolutional neural network that performs autoregressive audio waveform generation. By autoregressive, we mean that Wavenet predicts a conditional probability distribution sample given a sequence of past generated samples and selects the most probable discrete value from that distribution. The probability of the audio sequence is computed using the chain rule from the conditional probabilities of every individual sample given their previous samples, as follows:

$$p(x) = \prod_{t=1}^{T} p(x_t|x_1, x_2, ...x_{t-1}) \tag{2.10}$$

We have chosen WaveRNN as the vocoder to generate the sequence to sequence models evaluated in this thesis. WaveRNN is an autoregressive neural network like Wavenet that generates speech by sampling from a distribution. The network typically comprises of RNNs unlike Wavenet which is convolutional [Tachibana et al., 2018]. WaveNet has the problem that it works with up to 8 bit (μ-law) signals and requires large amounts of calculations. WaveRNN solves this problem. It works with up to 16 bit signal and requires fewer calculations than Wavenet [Amada et al., 2018]. The implementation used in our work is an adapted implementation of WaveRNN proposed in Tachibana et al. [2018] and the open-source implementation can be found in Fatchord [2019] which has the architecture shown in Figure 2.11. WaveRNN comprises ResNet which is a bank of recurrent neural networks, and an upsampling network. The FC Layers are fully connected feed-forward neural networks and the GRU layers are gated recurrent units which are similar to LSTMs (described in Section 2.1.2) in that they serve as short-term memory and have internal gate mechanisms that regulate the flow of information.

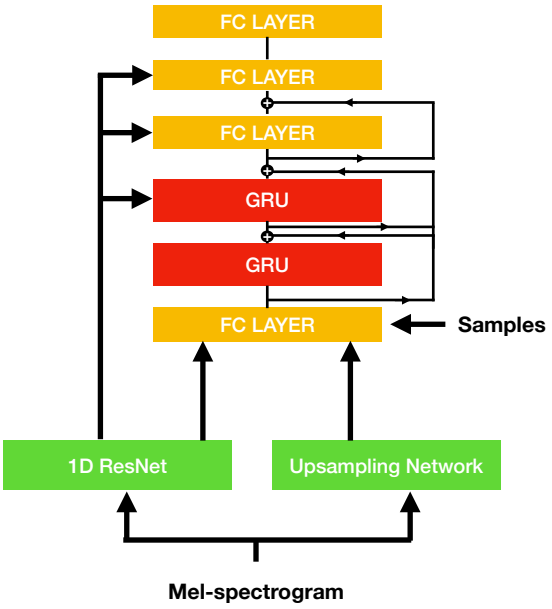


Figure 2.11: Diagram of the WaveRNN architecture

2.2 Perception of speech produced by TTS

Work on the perception of synthetic speech first began in the early 1970s where researchers were primarily interested in evaluating the segmental intelligibility of synthetic speech in comparison to natural speech. The main purpose of such studies was to close the intelligibility gap between natural and synthetic speech by identifying the segments of synthetic speech that needed improvement. The bulk of perception studies focused on techniques that evaluated acceptability, preference, naturalness

and usefulness as well diagnostic techniques that evaluated segmental intelligibility and prosody.

Some researchers instead chose to conduct speech perception studies to understand *why* people perceived synthetic speech differently than natural speech. In this way, focus was placed on answering questions like "Why is synthetic speech more difficult to understand than natural speech?" and "Does synthetic speech lack certain fundamental characteristics of natural speech which might be helpful to perception?" [Winters and Pisoni, 2004]. This led to a long line of research that attempted to answer these kinds of questions.

In the early 1980s, Nusbaum and Pisoni [1985] hypothesized that the way humans perceive natural speech is different from the way they perceive synthetic speech produced by rule-based systems. They hypothesized that synthetic speech is equivalent to noisy natural speech. In other words, synthetic speech is perceived in the same way that natural speech is perceived in the presence of noise. It was for this reason that they believed that natural speech and synthetic speech are not comparable during speech perception studies as they do not belong on the same continuum. As an alternative hypothesis, synthetic speech is considered to be a perceptually impoverished form of natural speech because it lacks acoustic variability found in natural speech. This alternative hypothesis contrasts the idea that synthetic speech does not exist on the same continuum as natural speech. Both these hypotheses were investigated and results suggested that the first hypothesis should be rejected. Differences in speech perception between natural and synthetic speech were found to be affected by the acoustic-phonetic structure of the speech signal. At the time of this study, synthetic speech did not contain as many prosodic, acoustic and phonetic cues as natural speech, and as a result speech perception studies concluded that synthetic speech demands an increased mental workload compared to natural speech [Duffy and Pisoni, 1992].

For example, Pisoni et al. [1985] conducted a lexical decision task and results showed that response latencies were significantly longer for synthetic speech produced by rule-based systems than natural speech in the tasks of recognizing words and non-words. When measuring the effects that encoding synthetic speech has on the working memory, using word recall, it was found that processing synthetic speech in the brain impaired recall more than processing natural speech. These findings showed that synthetic speech required more processing capacity and resources from the short term memory [Luce et al., 1983]. Requiring more processing capacity to process synthetic speech means that other concurrent cognitive processes are affected because there will be fewer cognitive resources available.

Similar results were obtained when Manous and Pisoni [1984] applied the gating paradigm. In the gating paradigm, listeners were expected to recognize a word as successive presentations of a target word was played at varying duration. For example, the first 50 ms of the word was played, followed by 100 ms of the same word until the entire word was heard. More time was needed to recognize a synthetic word as opposed to a natural word. It was concluded that synthetic speech produced by rule-based systems lack the redundancy in acoustic phonetic structure that exists in natural speech such as pausing, lengthening on prominent syllables, intonation, hyper/hypo articulation and therefore takes longer to recognize.

It is evident that synthetic speech produced by rule-based systems is perceived differently to natural speech and findings consistently imply that synthetic speech is more difficult to process than natural speech. In most cases, the contributing factor to increased cognitive load was due to intelligibility. However, Pisoni et al. [1987] evaluated the comprehension of synthetic speech by controlling intelligibility so that it could be determined whether increased effort was due to other factors beyond the segmental intelligibility of the speech signal. Listeners in this experiment experienced difficulty in understanding and comprehending synthetic speech but had no difficulty in their ability to correctly transcribe the sentences. This implied that there are other factors that play a role in speech comprehension beyond the ability of the listener to merely encode the message, factors such as prosody for example.

Paris et al. [2000] explored the effect of prosody on the processing of synthetic speech produced by rule-based systems. Their results suggest that prosody is a contributing factor in degraded performance when listening to synthetic speech compared to natural speech. This supports the notion that prosodic cues are essential for speech processing as they speed up processing time. It is believed that, when there is a lack of prosodic cues, the listener's attention is shifted towards a more shallow form of processing which focuses on superficial acoustic information as opposed to deeper linguistic analysis that is required for comprehension. This shift in attention has been found in a number of studies involving increased resource demands on working memory for both natural and synthetic speech [Luce, 1982, Paris et al., 1995, Ralston et al., 1991].

The work described thus far mostly evaluated synthetic speech generated from rule-based systems. Rule-based systems were first generation TTS systems that relied on hand-crafted rules to convert the textual information to speech. Nowadays, the generation of synthetic speech has been taken over by automatic data driven approaches such as unit selection [Hunt and Black, 1996, Merritt et al., 2016] and statistical parametric speech synthesis [Zen et al., 2009, 2013] as discussed in Section 2.1. There was a gap in research with respect to speech perception studies evaluating comprehension of synthetic speech up until the work of Wester et al. [2015] which, to my knowledge, was one of the first speech perception studies that evaluated synthetic speech produced by data-driven methods. They explored the temporal delay hypothesis on natural, vocoded and synthetic speech. Temporal delay occurs when there is a delay in a word onset which is typically caused by filled pauses such as uh and um. It was hypothesised that this type of delay aids the listener in recognizing the word that follows the filled pause more quickly. Their findings show that response times for synthetic speech generated by SPSS are much longer than response times for natural speech. They hypothesized that this is because it requires more cognitive effort to process synthetic speech than natural or vocoded speech. Wester et al. [2015] suggested that the work done by Pisoni et al. [1985] and other researcher's in the field of speech perception should be revisited with SPSS as perception of current TTS speech quality has probably changed compared to earlier approaches.

Some TTS systems that exist today, namely sequence-to-sequence models, are capable of generating speech that is considered to be hard to distinguish from human speech. However, such results are obtained through standard evaluation measures that will be described in the next section. There is little to no knowledge on whether these results translate to TTS demanding cognitive resources

that are equivalent with those demanded when listening to human speech. This was the driving motivator for the experiments conducted in this thesis.

2.3 Evaluation of text-to-speech synthesis

Evaluation is a critical step in the developmental pipeline of text-to-speech synthesis as it is the process that helps us understand the overall performance of the system. Conventionally, particular focus is placed on the output speech quality that is produced by the built systems. Ideally, high performance is associated with speech that is natural sounding, in other words, as close to human speech as possible and highly intelligible.

There are two broad types of TTS evaluations, those that are conducted objectively and others subjectively. Objective evaluations rely on mathematical approaches (eg., mel-cepstral distortion or root mean square error of F0) to determine how close synthetic speech is to natural speech whilst subjective evaluations rely on human perception. Since the core purpose of text-to-speech synthesis systems is to make it usable in real-world applications, it is more meaningful to conduct experiments measuring the synthesizers' performance with human listeners and therefore subjective evaluations are more widely used. Therefore, in this thesis we focus more on subjective evaluations and only relevant tests adopted in this thesis are discussed.

The most common test for intelligibility is sentence recognition. Listeners are asked to listen to sentences and transcribe what they hear. The transcribed sentence of the listener is compared against the original sentence and a word error rate (WER) is computed using the following calculation:

$$\text{WER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{N}} \times 100 \tag{2.11}$$

where N is the total number of words in the sentence.

Some researchers perform this test using Harvard sentences [Rothauser, 1969] which contain a natural distribution of phonemes in English. However, when using such sentences, listeners can guess words which can affect the accuracy of the results. Therefore, a more common approach is to use semantically unpredictable sentences (SUS) [Benoît et al., 1996]. These sentences follow the syntactical "rules" of English but lack any form of a predictable meaning which reduces the possibility of listeners guessing the next word. Although the results obtained using SUS are more accurate, one can argue that using such sentences that do not exist in real-world have little validity in terms of a real world applications. Therefore, sentence material is an important design choice when performing evaluations.

To evaluate the naturalness of synthetic speech, listeners are played speech samples and are asked to rate how natural the speech sounds. These ratings are often in the form of a mean opinion score (MOS), where listeners rate their opinions on a 5-point Likert scale starting from 1 (completely unnatural) to 5 (completely natural). Such tests are biased towards what individual listeners perceive as being natural and this is something that isn't easy to normalize across listeners. If all listeners have a golden standard to compare against then human speech would always be rated

as 5, but this isn't always obtained. There are other influences like pleasantness and likability of a given voice that influence human perception of naturalness. Furthermore, the scores obtained in MOS tests are also not comparable to scores of another MOS test and therefore it is well known that such a test can only be used as a form of ranking [Taylor, 2009]. Additionally, Tokuda and Black [2005] pointed out that for this type of tests, it is more correct to report the MOS score than their respective mean values and so this is the approach followed in this thesis.

The Blizzard Challenge was introduced in 2005 [Tokuda and Black, 2005] and its purpose was to extensively evaluate TTS on an internationally large scale. TTS developers could submit their TTS systems that were trained using a standardised training set provided by the organisers of the challenge and as a result meaningful comparisons to other systems could be obtained. The systems and results of some speech synthesizers evaluated in the 2010 and 2011 Blizzard Challenges were used as a starting point in this thesis. Comparing results of systems that have already been previously evaluated with the results obtained from our proposed evaluation methods gives us an indication of the accuracy achieved in our proposed methods. It is important to validate our results against older methods before using the new method to evaluate new TTS systems.

King et al. [2017] shows that some synthesizers have been given high ratings of naturalness and achieve close to ceiling intelligibility scores. High ratings indicate that the voices are approaching close-to human sounding speech. Once TTS voices are rated 5 in naturalness on a 5-point MOS Likert scale (which is completely natural) and obtain close to 100% intelligibility levels, current evaluation methods will eventually become meaningless and outdated. Achieving close-to human-like sounding speech means TTS will become increasingly used in many real-world applications and therefore evaluation methods need to change direction. Emphasis in evaluation metrics needs to be placed more on the impacts it has on the end-user of the real-world applications embedding TTS. As discussed in Section 2.2, it has been consistently shown in the past that synthetic speech imposes greater cognitive load than natural speech. Therefore, understanding whether this theory holds true requires investigation. There is a lack of research with regards to measuring cognitive load imposed by state-of-the-art text-to-speech systems and thus there is little understanding of how modern synthetic speech interacts with the human cognitive processing system. Thus far, little to no attempt has been made in understanding the negative impact that current TTS systems potentially have on the listener. In Part I of this thesis we investigate methods for measuring the cognitive load of synthetic speech.

Furthermore, most listening tests are conducted in ideal (quiet) listening conditions, yet end-users in a realistic scenario won't always be listening to TTS under such conditions. It is well understood that listening in the presence of noise is harder than listening in quiet and so evaluation methods also need to move towards more realistic evaluation strategies. Simantiraki et al. [2018] evaluated TTS in the presence of speech-shaped noise and found that HMM-based TTS was the most difficult to listen to. When listeners hear speech they automatically match the rapid incoming acoustic stream to stored representations of words and phonemes in memory to successfully extract the intended meaning [Peelle et al., 2010]. The process of correctly identifying sounds is made more difficult when speech is acoustically degraded, like synthetic speech, and thus less information is

available to the listener. As a consequence, quality of speech cues is reduced and the effort required to listen becomes greater. The question this raises is: what factors in synthetic speech are resulting in this hard-to-listen-to effect, or an increase in effort required? In Part II of this thesis, we set out to investigate what are the contributing factors that lead to an increased cognitive load - if any - in both quiet and noisy conditions.

2.4 Cognitive load

Cognitive load (CL) refers to the amount of pressure placed on a person's working memory when performing a task. In the field of speech perception and understanding, the term often used to describe cognitive load is listening effort (LE). LE is defined as the deliberate allocation of mental resources to the task of processing speech and thus increased listening effort means more brain power is needed to recognize and understand speech. Research investigating listening effort identified the dominant cognitive processes for speech processing as **working memory**, **attention**, **processing speed** and **linguistic knowledge** [Gagne et al., 2017]. Whilst the scope of this thesis is not to understand each cognitive process individually, these four processes are vital for understanding results presented later in this thesis. We will discuss these processes in more detail as they arise in the discussion of the results.

CL can be measured in various ways and the type of measurement typically falls under one of three broad categories: behavioural, physiological and self-reporting assessments. Behavioural measurements include the dual-task [Picou and Ricketts, 2014, Seeman and Sims, 2015, Gagne et al., 2017] and visual word paradigms [Klingner et al., 2011, Picou et al., 2011]. Physiological measurements include methods such as pupillometry [Zekveld et al., 2018, Koelewijn et al., 2012], skin conductance [Mackersie and Calderon-Moultrie, 2016], electroencephalography (EEG) [Bernarding et al., 2012, Damian et al., 2015], functional magnetic resonance imaging (fMRI) [Peelle et al., 2010, Peelle, 2018] and functional fear-Infrared (fNIR) [Ayaz et al., 2013]. Self-reporting measures include questionnaires, task load index or categorical rating scales [Alhanbali et al., 2017]. In this thesis, one assessment from each of the three broad categories was chosen to measure the cognitive load of synthetic speech.

Behavioural measurements In the category of behavioural measurements, the dual-task paradigm is the most common assessment. The dual-task paradigm is based on the assumption that human cognitive capacity is limited. Therefore, when two tasks are performed concurrently and one task, for example the primary task, is more cognitively demanding than the other task (secondary task), this will result in reduced performance on the secondary task if the primary task is prioritized. Under the limited capacity assumption, eventually all available mental resources will be utilized and as a consequence performance in the secondary task deteriorates. In speech understanding studies where the primary task involves listening to speech, researchers have shown that performance on a secondary task can change as a function of changes in speech quality [Gosselin and Gagné, 2011, Sarampalis et al., 2009]. Therefore, cognitive load is measured as the difference in performance

between performing the secondary task in isolation and the performance of the same task in the dual-task. Therefore, this type of behavioral measurement is an indirect test of the cognitive load required to process speech. Since the dual-task paradigm is the most common behavioral method and has extensive supporting material in the literature, it was the chosen method for measuring the cognitive load of synthetic speech in our experiments. A detailed survey of work applying the dual-task paradigm specifically for the purpose of speech perception is provided in the next chapter.

Physiological measurements Physiological measurements of cognitive load are those that rely on bodily responses that occur when the cognitive processing system is placed under strain when performing a task. The advantage of such a measurement is that these responses take place without the listener being aware of them and therefore are not directly influenced by subjective biases. The most common physiological methods for measuring cognitive load are neuro-imaging methods like fMRI and fNIR or other methods like pupillometry. In neuro-imaging studies, brain activations are captured when a participant carries out a cognitive task. For example in a speech-related study, Benson et al. [2001] captured the brain activation using fMRI of participants whilst listening to both natural and synthetic speech. The study demonstrated that brain activation changes with respect to the quality of speech when using fMRI. As speech degraded in quality, a higher activation was observed. This finding suggested that additional cognitive processing takes place to help the brain deal with modulation in speech quality to keep a high performance in the task of comprehension. The same authors conducted a study that identified neural correlates of speech processing using functional near infrared (fNIR) spectroscopy. fNIR measures blood oxygenation and blood volume during the performance of a cognitive task. Oxygenation changes were used as the objective measure for cognitive load. The results showed that natural speech was associated with significantly lower oxygenation compared to synthetic speech. Lower oxygenation levels indicated minimum effort. In contrast, synthetic speech showed higher oxygenation levels suggesting increased effort. The limitation of such approaches is that neuro-imaging requires special equipment that is expensive and not easily accessible.

Pupillometry is another common physiological method that has gained much traction over recent years and is non-intrusive. Pupillometry requires only an eye-tracker which is typically inexpensive in comparison to neuro-imaging equipment. Pupillometry, like neuro-imaging, is an online measure. It tracks the pupil response whilst performing a cognitive task. The size of the pupil has been consistently shown to be positively correlated with cognitive effort and task difficulty [Beatty, 1982, Zekveld et al., 2011, Koelewijn et al., 2012, Kramer et al., 2013, Kuchinsky et al., 2013]. The more effort exerted, the more the pupil dilates. On the basis of easy accessibility to equipment, pupillometry was the physiological method chosen in this thesis and therefore a more detailed literature review is provided in a later chapter.

Self-reporting measures This type of measurement gives a subjective rating of cognitive load which is typically assessed in combination with behavioural or physiological measurements as they are usually used to support the results found in the behavioural and physiological measurements.

Typical self-reporting measures include questions the listener is asked such as "How much effort did you exert when listening?" or "How tired did you feel whilst listening?" and their responses are recorded typically using a Likert rating scale. In some studies, the self-reporting measures correlate with the objective measures, whilst in most other studies they don't [Sarampalis et al., 2009]. In this thesis, three 5-point rating scales were used to collect self-reported cognitive load, listeners' perception of naturalness.

Part I

Measuring cognitive load of synthetic speech

Chapter 3

The dual-task paradigm

In this chapter, the dual-task paradigm is explored as a potential method for measuring cognitive load of synthetic speech. The aim of these experiments is to answer the research question (RQ1) on whether such a method is sensitive to differences in cognitive load between various speech synthesizers and human speech.

We start this chapter by describing how the dual-task paradigm works, followed by outlining key factors influencing listening effort reported by investigators who have employed the dual-task paradigm. The key factors discussed are those that were considered during the experimental design of the paradigm presented in our work. We then describe our methodology and implementation of the dual-task paradigm. This is followed by the presentation of our results and we conclude this chapter by summarizing the key findings.

(This chapter expands Govender and King [2018a])

3.1 Introduction

When processing speech in a realistic scenario, we are often required to multi-task. For example, when we wish to listen to an individual in a crowded room, we often have distracting activities taking place around us. Therefore, whilst listening we are required to focus our attention only on that individual we wish to listen to. By prioritising the listening task we hope to block out those distractions that exist around us in order to improve speech understanding. The dual-task paradigm attempts to simulate multi-tasking in a similar manner that expects a listener to listen to speech whilst simultaneously performing a distracting task.

The idea behind the dual-task paradigm is based on a hypothesis that human cognitive capacity is limited [Sweller, 2011]. In this hypothesis, when the brain is forced to process two tasks simultaneously, the total mental resources available would need to be split between those two tasks. However, if one task is prioritised over the other, then more resources are *deliberately* allocated to the prioritised (primary) task which naturally increases task performance. Using prioritization, we thus have control over shifting the allocation of our mental resources. As a consequence, fewer resources remain to process the other (secondary) task which leads to a deterioration in performance

on the secondary task. Now, depending on how demanding the primary task is, performance in the secondary task will decline accordingly. In other words, a more cognitively demanding primary task will cause poorer performance in the secondary task. This decline in performance is almost guaranteed if the total amount of resources needed for both tasks exceeds the total human capacity [Gagne et al., 2017]. Thus, the amount of mental effort exerted in a given primary task is calculated as the difference in performance on a secondary task between performing the task in isolation (baseline) and in a dual-task. This difference is referred to as the dual-task cost. It is important to realise that this hypothesis only holds true if the performance of the primary task remains the same in both baseline and dual-task conditions.

3.2 Methodology

Gagne et al. [2017] presented a review of previous experiments employing the dual-task paradigm in relation to listening effort for speech understanding. This review was used as a guideline for designing the dual-task paradigm applied in this chapter. The key findings presented in this review highlighted the main factors that potentially influence the listening effort measurement when employing the dual-task paradigm. Each of these key factors are described in the sections that follow.

3.2.1 Target Listeners

Age and hearing status are key factors such that older adults expend greater listening effort than younger adults under the same listening conditions in noise [Gosselin and Gagné, 2011, Helfer et al., 2010] and listeners with hearing loss exert more listening effort than their age-matched counterparts with normal hearing acuity [Desjardins and Doherty, 2014, Neher et al., 2014]. For these reasons, in our experiments we only considered normal-hearing young adult listeners typically in the range of 18 to 30 years old.

3.2.2 Modality

The modality in which stimuli are presented to listeners was shown to have contrasting findings in two studies [Gosselin and Gagné, 2011, Picou and Ricketts, 2014]. The first study claimed that it is more effortful to listen in an auditory-alone mode than one with the provision of audio-visuals whilst the second study found no significant differences. Although these findings contradict each other, there was also a difference with respect to the manner in which these paradigms were applied. The former was administered in a parallel experimental setup whilst the latter followed a sequential setup. Amongst other works reviewed in Gagne et al. [2017], the majority of the studies applied a parallel experimental paradigm. According to the authors, the parallel approach is one that simulates a real-life multi-tasking scenario and therefore this is considered to be more ecologically valid. Also, in order to apply the sequential approach, the tasks involved typically requires some form of memorisation and therefore it is said that parallel dual-tasking taps more into utilizing processing resources rather than memory resources. To date it remains unknown whether or not these two

setups measure the same dimensions of listening effort. Since ecological validity is an important factor and given that text-to-speech synthesis deployed in real-world applications is mostly listened to in auditory-only modality, we chose to follow a parallel experimental setup in this work.

3.2.3 Cognitive Abilities

With respect to the relationship between dual-task cost and an individual's cognitive abilities, results presented in work so far were found to be inconclusive [Gagne et al., 2017]. Therefore this factor was not a major concern in our experiments, but to some degree it was controlled given that all participants recruited were university students and therefore we assume that the variability of cognitive abilities amongst these candidates do not differ significantly.

3.2.4 Self-reported measures

The relationship between behavioral and self-reported measures (as defined in Section 2.4) of listening effort have also been previously investigated [Gosselin and Gagné, 2011, Fraser et al., 2010]. In general, no associations between the two types of measures were observed. This indicates that the dual-task paradigm is probably not indexing the same attributes of listening effort that listeners use when asked to rate their listening effort. Therefore listening effort measured by behavioural experiments is often indexing something different to what is perceived as effort by the listener. As mentioned in Section 2.4, self-reporting measures are often collected in support of behavioural measurements and therefore we collected both behavioural and self-reporting measures in our experiments. Although previous work found no associations between self-reported measures and behavioural measures, it was still important for us to collect them in this work. We believe that such findings can still be meaningful and provide valuable insights or raise important questions such as the validity of the self-reported measurements or discrepancies due to subjective biases that are absent in behavioural measurements. Thus, negative implications of observing the same result is minimal and can instead be utilised to better understand the validity of the results.

3.2.5 Task Selection

There are two key factors that are difficult to control yet critical for gaining accurate results in the dual-task paradigm: motivation and difficulty. It is important that the listener remains motivated at high levels for both the primary and secondary tasks. Selecting a task that is difficult enough is vital such that the amount of processing resources that are required exceeds the resources that are available. In this way, deterioration in the secondary task is almost certainly guaranteed. If the task is not challenging enough, one will not know whether a statistically insignificant result is due to no difference in effort or due to both primary and secondary tasks being performed comfortably with the total resources available, resulting in no compromised performance on the secondary task. Therefore, great consideration needs to be taken when choosing the task for this paradigm.

Speech recognition with recall of either sentences, words or syllables is the most common primary task used [Wild et al., 2012, Hornsby, 2013, Picou et al., 2016]. For the secondary task, various

types of tasks have been tried such as visual probe tasks [Picou and Ricketts, 2014], auditory distractors [Wild et al., 2012] and visual motor tracking tasks [Desjardins and Doherty, 2014]. In most previous work, the choice of secondary task was rarely motivated. Therefore, a main concern with the dual-task paradigm is the wide variability in experimental designs for measuring listening effort, especially in the choice of secondary task [Gagne et al., 2017]. On one hand, Wu et al. [2014] compared two secondary tasks and in both cases a correlation to listening effort was found. This supports the idea that different secondary tasks could be equally appropriate and therefore shouldn't affect the outcome of the experiment. On the other hand, the effect of changing the secondary task by comparing a simple visual probe, a complex visual probe and a word-category recognition task were investigated by Picou and Ricketts [2014] and results showed that only the word-category recognition task affected performance on the primary task. The investigators believe that the word-category recognition task required deeper processing than the two visual-probe tasks. However, they were uncertain whether it was more sensitive because both primary and secondary tasks were linguistic or because the word-category recognition task was simply more demanding than the visual probes. This clearly indicates that it is still unclear whether *any* secondary task can be used to investigate listening effort. Further research is therefore required to identify which type of secondary task or combination of primary and secondary tasks are best suited for investigating listening effort for speech understanding [Gagne et al., 2017]. Despite previous work not identifying a suitable secondary task, it is important to note that majority of these studies evaluated the listening effort of human speech. Therefore, there is a possibility that listening effort demanded by human speech is more difficult to detect than synthetic speech. Since previous studies (not only those using the dual-task paradigm) have all consistently shown that synthetic speech demands more cognitive load than human speech, the motivation behind investigating whether the dual-task paradigm is sensitive to detecting cognitive load differences between various speech synthesizers was still unknown.

In [Gagne et al., 2017], most primary tasks listed involved a word recognition task, or in the instance that sentences were used only the recognition of specific keywords were measured. In our experiments, we chose to use sentence recognition as the primary task as this is the approach used when evaluating text-to-speech synthesis in conventional listening tests. The difference between previous primary tasks and ours is that we expected the listener to recognize all words in the sentence as is expected in traditional TTS listening tests. For the secondary task, we explored two types of a visual-motor task: a digit task and word task. We chose to investigate two types of visual probe tasks in case the digit task was found to not be cognitively demanding enough. We hypothesised that by using a secondary task that taps into linguistic mental resources, given that the primary task taps into linguistic mental resources as well, then perhaps a greater overall load on the user will be imposed. Therefore, increasing the likelihood of observing deterioration in the secondary task. The digit and word task are described in more detail later.

3.2.6 Task Priority

Another methodological concern is that priority should be given to the primary task by the listener under all circumstances for the limited cognitive capacity hypothesis to hold true but this

is of course something that isn't always easy to measure. In our experiments, this was addressed by instructing the listeners to verbally recall the sentences they heard which were recorded by a microphone. For each participant, these recordings were informally checked against the sentences they were required to listen to.

3.2.7 Application to synthetic speech

To our knowledge, Sonntag et al. [1998] is the only study that used the dual-task paradigm to measure listening effort of synthetic speech. The investigators attempted to simulate a real-life scenario of a driver listening to a navigation system. The primary task was an auditory-verbal task and the secondary task was a visual-motor task. The primary task was a simple calculation involving addition and subtraction and the secondary task required the participant to select one out of four colour keys which matched the color that appeared on a monitor. All stimuli presented during the task differed in meaning but the syntactic structure was always the same. In this way no world knowledge was required and therefore comprehension effort was not influenced by individual differences. Accuracy and response times were measured. Significant differences were found only between the human voice and all synthesizers but no significant differences were found between the various speech synthesizers. The most significant difference was between the human voice and poorest quality synthesizer.

A major limitation of this study is that the synthesizers compared were trained using different speech material including speech recorded by both males and females. Although the result tells us that synthetic speech is more effortful to listen to than human speech, by not controlling the material used to train the synthesizers, it makes it difficult to pinpoint why this is the case. Whilst it is important to design a paradigm that has the ability to evaluate the listening effort of synthetic speech, it is also important to take careful consideration in the design choices of the paradigm such that we can maximize the insights gained from the results. Being able to understand why synthetic speech has increased cognitive load compared to human speech becomes more important as this technology is employed in real-world applications. Understanding *why* provides us with the necessary information required to suggest better ways to develop speech synthesizers that impose low cognitive load on the end-users. Since 2005, evaluating speech synthesizers with controlled speech data was made possible through the Blizzard Challenge [Tokuda and Black, 2005]. By eliminating the variability of the data, much fairer comparisons across speech synthesizers can now be obtained. By using controlled speech data, more meaningful results from the dual-task paradigm could be obtained than in [Sonntag et al., 1998].

The experiments carried out in this chapter addresses some of the problems recognized above by comparing speech synthesizers from a single Blizzard Challenge together with the corresponding human speech that was used to build them. Our hypothesis is that by comparing TTS systems whereby the speech data and sentence material is controlled will provide a fairer comparison across the speech synthesizers which will enable us to detect differences in cognitive load between the various speech synthesizers compared based only on their architectural differences.

3.3 Implementation

3.3.1 Tasks

The primary task used was a sentence recognition task. The listener was instructed to listen to a spoken audio sample and verbally repeat the sentence as accurately as possible. The secondary task was a visual motor task. Two types of a visual motor task was investigated in our work: a digit task and word task. The dual-task paradigm was administered in three experimental conditions: (1) Secondary-task alone (Digit/Word Task) and (2) Practice Dual and (3) Dual-task, set-up as follows:

Digit Task In this secondary task, the listener needed to decide if a visually-displayed digit is odd or even. A single trial consisted of two digits displayed sequentially on the screen. The listener needed to respond as quickly as possible by pressing a button on the response box placed in front of them. For every trial, the first number appeared at a random delay time after the onset of the trial. After the first response was collected, the second digit immediately appeared. Only participants with accuracy greater than 85% were allowed to proceed to the next task condition. A criteria of 85% accuracy was applied as this was an indication that participants reached close to ceiling performance in the secondary-task alone and therefore would minimize the training effect[1] from occurring during the dual-task. Figure 3.1 illustrates the digit task deployed in this thesis.

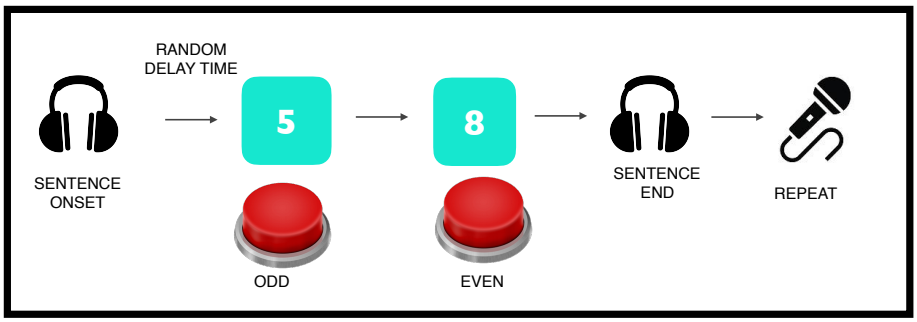


Figure 3.1: Illustration of secondary digit task in our dual-task paradigm

Word Task This secondary task was inspired by Picou et al. [2016] who suggested that a word-category secondary task is sensitive to listening effort. The motivation is that, given that both tasks are linguistic, the secondary task requires the listener to make use of the same (limited) cognitive resources as the primary task, which should result in a greater cognitive load. It is performed exactly like the digit task (described above) except the listener needed to decide if a visually-displayed word exists or not. Common 4 letter words such as "home" or "boat" were used to minimize the likelihood that a listener had not heard the word before which would put them at a disadvantage when performing the task. Figure 3.2 illustrates the word task deployed in this thesis.

[1] A training effect occurs when a participant gets better at performing a given task with increased exposure. Therefore, in the dual-task this would mean that as the experiment progresses, participants reaction times could get faster

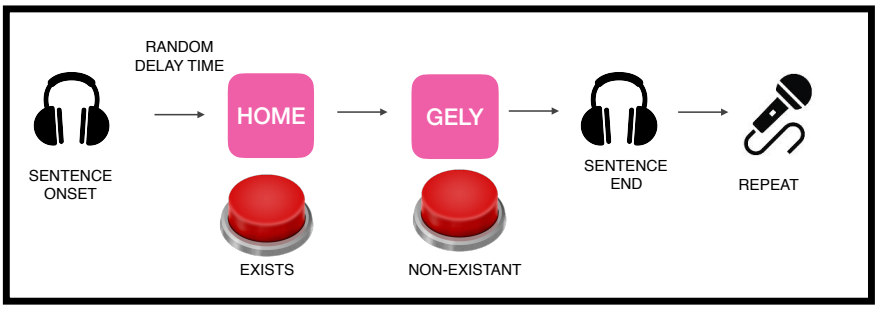


Figure 3.2: Illustration of secondary word task in our dual-task paradigm

Practice dual The goal of the Practice dual was to allow the participant to familiarize him/herself with the dual-task. The practice dual consisted of three trials. Both the primary and secondary task (Digit/Word) were performed concurrently. The audio in this task consisted only of human speech to avoid exposure to the speech conditions in the main dual-task that followed. The participants were allowed to perform the practice dual as many times as they wished until they felt confident to proceed to the final dual-task. This was motivated by the desire to reduce the training effect that could take place during the main dual-task experiment.

Dual-task The dual-task was set up in the same way as the digit/word task but now included the listening task (primary task). Whilst the digits/words were displayed, the participant was expected to listen to an audio sample concurrently. They were instructed to prioritize the listening task as they were expected to recall the sentence at the end of each trial. They were told to respond to the visually-displayed digit/word as fast as they could and repeat the sentences verbatim. Their verbal responses were recorded in order to confirm that they were prioritising the primary task. At the end of each trial, self-reporting measures were collected. Each participant was asked to rate the naturalness of each audio sample and the difficulty of listening to it, each on a 5-point scale labelled 1 - very natural to 5 - very unnatural and 1 - very easy to 5 - very difficult respectively. Each audio sample was played diotically to the listeners using Beyerdynamic DT770 headphones in individual soundproof booths. Stimuli were presented using E-Prime 2.0 software [Schneider and Zuccolotto, 2012]. Reaction times for the secondary task and the self-reported responses were recorded with an E-Prime response box and their verbal responses were recorded using a microphone.

3.3.2 Structure

The dual-task paradigm was presented to the participant in 11 blocks. The first block was regarded as a familiarization block and results were discarded. In the familiarization block, a sentence from each evaluated speech condition in Table 3.1 was heard. The ten remaining blocks consisted of 10 sentences each and were constructed using a 5 x 5 Latin square to balance all listeners and speech conditions, repeated twice. Each of these blocks used five audio samples all from only one of the five conditions, with each condition appearing in two blocks. All audio samples on average contained 8 words and were approximately 2.5 to 3 seconds in duration.

Table 3.1: Summary of selected speech synthesis systems, with their scores from the Blizzard Challenge 2011 for naturalness (Median Opinion Score, MOS – higher is better) and intelligibility (Word Error Rate, WER – lower is better)

Speech Condition	**Naturalness (MOS Score)**	**Intelligibility (WER%)**
Natural (Human)	5	16
Hybrid	3	20
Unit Selection	3	22
HMM	3	20
Low Quality HMM	1	26

3.3.3 Stimuli and Sentence Material

Stimuli presented to the participants were sentences generated by four speech synthesizers taken from the 2011 Blizzard Challenge and the human voice used to build them [King and Karaiskos, 2011]. These sentences can be found in Appendix A. All synthesizers were built using 16.6 hours of speech data from an American English female professional voice talent. This dataset was chosen as it is the latest Blizzard challenge that contained natural and synthesized versions of both predictable and unpredictable sentence material necessary for the investigations carried out in this chapter. A summary of the selected speech synthesizers together with the scores collected during the 2011 Blizzard Challenge can be found in Table 3.1. Performance in terms of Mean-Opinion-Score (MOS) and Word-error-rate (WER%) for naturalness and intelligibility are shown respectively.

In conventional listening tests, intelligibility of synthetic speech is measured using semantically unpredictable sentences (SUS) [Benoît et al., 1996]. This is to ensure that participants are not using any linguistic cues to guess what they heard and forces them to respond only with the information they have heard. Therefore, all test sentences used in this work were semantically unpredictable and were not included in the training data. An example of a SUS sentences used is, "The old shape attacked the shoe." One might argue that SUS sentences are not ecologically valid and may influence the listening effort measurement. However, as mentioned earlier, differences in listener performance on the secondary task will only be observed if the total cognitive load on the listener exceeds capacity. Since all audio samples used in this experiment were considered to be clean speech, it was a concern that this would impose insufficient load on the listener because listening to human speech in quiet conditions is generally effortless [McGarrigle et al., 2014]. Johnsrude and Rodd [2016] explains that increasing demands placed on processing can be a result of several factors including linguistic properties of the stimulus like understanding a sentence that is syntactically correct but semantically incorrect. Therefore we opted to use SUS in the primary task to increase load.

3.4 Experiments

Two experiments were carried out. Experiment 1 employed the dual-task with the digit task as the secondary task and Experiment 2 employed the word task instead of the digit task.

3.4.1 Participants

20 students from the University of Edinburgh were recruited for each experiment. The participants' ages ranged from 18 to 28 years old. All participants were paid for their participation (£6). They were all native English speakers and were expected to fill out a consent form that asked them if they have any hearing problems. No participants reported any hearing problems. Ethical approval was obtained from the Informatics Ethics Committee.

3.4.2 Analysis

The measure of listening effort in our experiments is the difference in reaction time (RT) between the baseline and the dual-task on the secondary task. To exclude the effect of inter-subject differences, the proportional dual cost time (pDCT) was computed using the following equation (from Gagne et al. [2017]):

$$pDCT = \frac{RT_D - RT_S}{RT_S} \times 100 \tag{3.1}$$

where RT_S is the mean RT for the secondary-alone task and RT_D is the mean RT for the dual-task condition.

If the listener optimizes their performance on the primary task, it is assumed that the listener will perform equally well for both conditions on the primary task. In the dual-task, listening effort is measured using only trials where participants responded correctly. RTs greater than 2 s or less than 100 ms were marked as outliers and discarded. Responding as quick as 100 ms and slower than 2 s are likely to be response errors since the response time was limited to a 2 s interval and thus any time after this would be invalid.

Analysis of variance (ANOVA) with repeated measures was computed to infer statistical significance of results. The null hypothesis was rejected for p-values less than 0.05. Tukey's post-hoc test was subsequently applied to find all statistically significant pairs. To analyse scores collected in the self-reports, the Wilcoxon signed rank test was used to compute significance for each speech condition evaluated. Furthermore, the Pearson correlation test was used to determine the correlation between the self-reported naturalness scores and cognitive load.

We hypothesised that we would observe differences in cognitive load between human speech and synthetic in the digit task but not between the various speech synthesizers. However, in the word task we expected to observe difference between human speech and synthetic speech as well as between each speech synthesizer. Our hypothesis in terms of CL was that human speech would demand the least cognitive load followed by Hybrid speech, Unit Selection, HMM and Low-Quality HMM would demand the most.

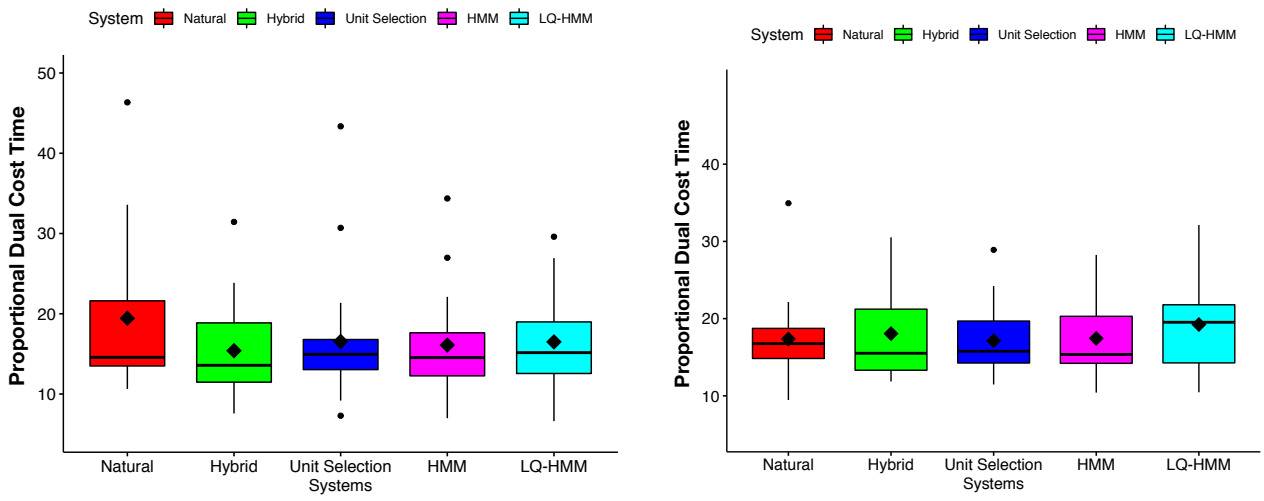


Figure 3.3: Boxplots showing the pDCT when listening to each speech condition in the dual-task paradigm, (a) Exp 1: Digit task (b) Exp 2: Word task

3.5 Results

3.5.1 Reaction Time

In each experiment 2000 reaction times were collected. 93% and 94% of trials were correctly answered in the dual-task condition for Exp. 1 and Exp. 2 respectively. In both experiments, it was observed that most of the incorrect responses during the dual-task condition occurred when participants were listening to the HMM and Low-Quality HMM (LQ-HMM) synthesizers.

In the final analysis, 88% (Exp 1) and 89% (Exp 2) of RTs remained after the exclusion of outliers which were taken as 1.5 interquartile ranges (IQRs) below the first quartile or above the third quartile. The differences between RTs in the dual-task and secondary-task alone was computed. The distribution of these pDCTs for both experiments are shown in Figure 3.3.

The analysis of variance with repeated measures indicated that the speech condition has a significant effect on the pDCT in Exp.1, where $F(4,76) = 2.74$ with $p=0.03$ ($p \leq 0.05$). A post-hoc Tukey test was then applied to identify significant pairs. Significance was only found between Natural speech and the Hybrid speech synthesizer ($p \approx 0.03$). This result was unexpected, as this results shows that RT is slower when listening to Natural speech compared to Hybrid synthetic speech. If this paradigm is measuring listening effort, this would suggest that synthetic speech is easier to listen to than human speech, which seems unlikely.

Upon further investigation, we observed that 50% of participants responded faster in the dual-task than they did in the secondary-alone task. Therefore, deterioration in the secondary task during the dual condition was not always achieved. As hypothesised, this led us to believe that both the primary and secondary digit task were not cognitively demanding enough. In other words, both the primary task and secondary digit task were manageable to perform concurrently without exerting a

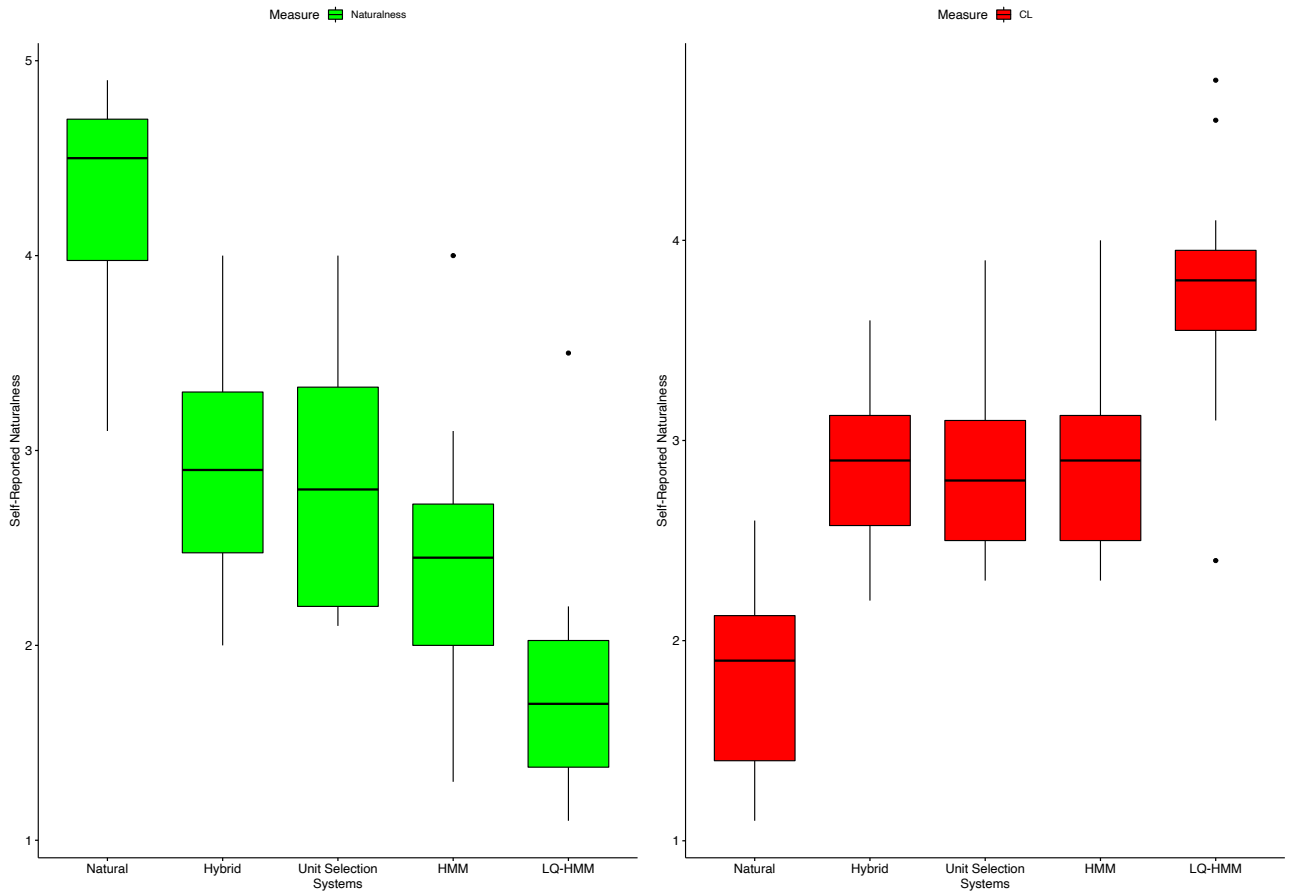


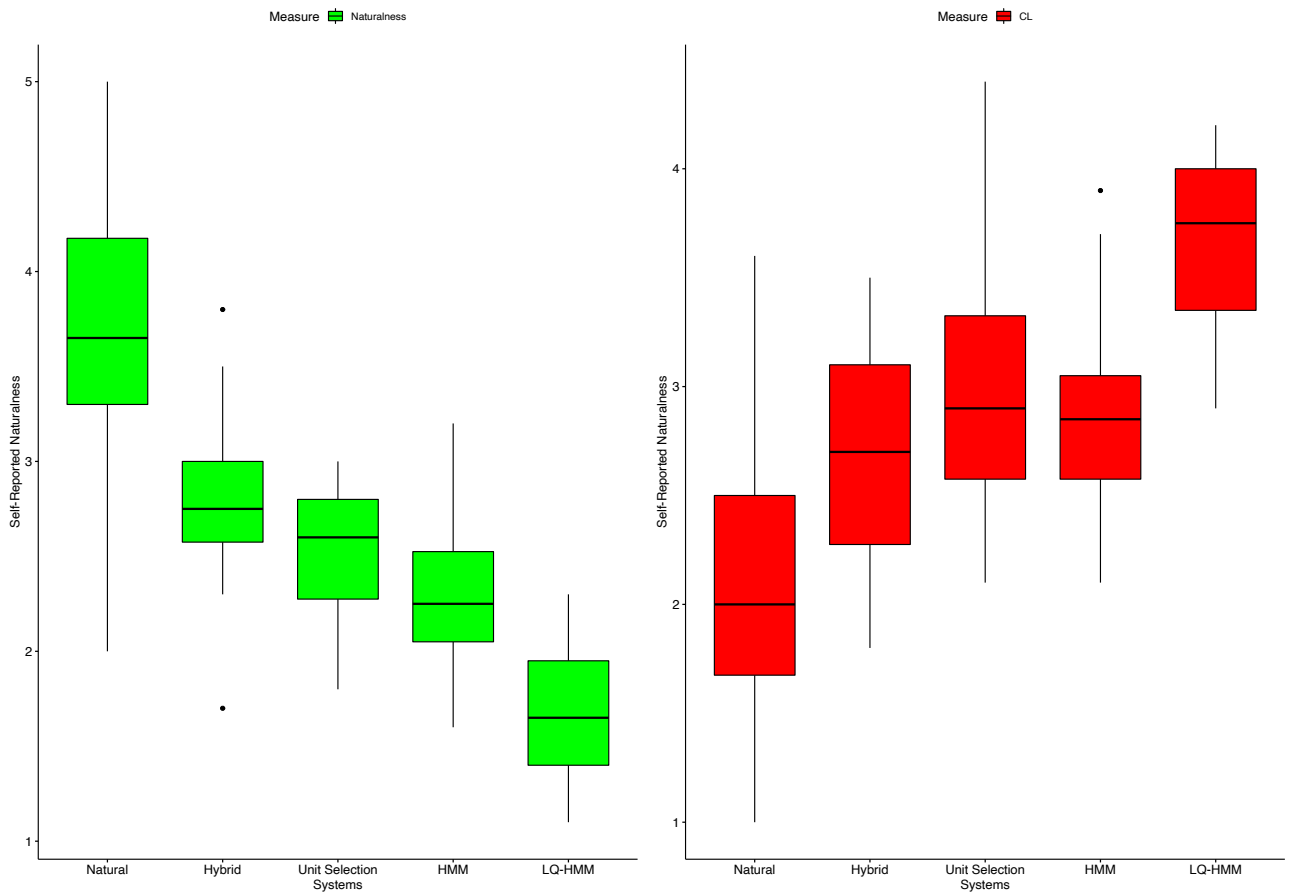
Figure 3.4: Boxplots showing the self-reported measures for Exp. 1 reported by participants on 5-point Likert rating scales for naturalness and cognitive load labelled 1 - very unnatural to 5 - very natural and 1 - very easy to 5 - very difficult, (a) Naturalness and (b) Cognitive load

forceful deterioration on the secondary task. As a consequence, it is unclear whether these results are in fact a real indication of listening effort as it is likely that both the primary and secondary digit task were performed comfortably with the total resources available, resulting in no compromised performance on the secondary task.

In Exp. 2, differences in pDCTs between speech conditions were statistically non-significant where $F(4,76)=1.29$ with $p=0.28$ ($p \geq 0.05$). The pDCTs in Exp. 2 were only marginally slower than Exp. 1 for all speech synthesizers which implies that the load demanded by the digit task and word task did not significantly differ and was not what we expected to find.

3.5.2 Self-reported measures

The self-reported measures are presented in Figure 3.5. In both experiments, listeners found natural speech to be the most natural sounding whilst the Low-Quality HMM was the most unnatural. Hybrid and Unit Selection were scored the same in both cases even though we expected Hybrid to be scored higher. Comparing the naturalness MOS in our experiments with those in the Blizzard Challenge shown in Table 3.1, the absolute median values have shifted slightly but the ranking of the systems from best to worst remained consistent with our predictions.



(a) (b)

Figure 3.5: Boxplots showing the self-reported measures for Exp. 2 reported by participants on 5-point Likert rating scales for naturalness and cognitive load labelled 1 - very unnatural to 5 - very natural and 1 - very easy to 5 - very difficult, (a) Naturalness and (b) Cognitive load

The self-reported cognitive load appears to be negatively correlated with naturalness and shows that natural speech is the easiest to listen to whilst the Low-Quality HMM was the hardest to listen to. Both Hybrid and Unit Selection scored the same in both cases. In Exp. 1, using the Wilcoxon signed rank test, all speech conditions compared were found to be statistically different ($p \leq 0.05$) in naturalness except the Hybrid and Unit Selection system. Similarly for cognitive load only the Hybrid, HMM and Unit Selection systems were found to be equivalent. In Exp. 2, all speech conditions were statistically different in both naturalness and cognitive load except for the Unit Selection and HMM systems. These results indicate, that when using the word task as a secondary task, the Hybrid system was perceived to be easiest to listen to compared to all other synthesizers yet during the digit task it was found to be equivalent to Unit Selection and the HMM system. By increasing the load on the secondary task, the more natural speech synthesizer was perceptually teased apart from the rest. This tells us that it is possible to detect differences between the speech synthesizer when the load is great enough.

Furthermore, by investigating the correlation between the two self-reported measures, a strong negative correlation was found in both experiments (Exp. 1: corr=-0.73 and Exp. 2: corr=-0.76 with $p \leq 0.05$). It is not surprising that the self-reported cognitive load is strongly negatively correlated with naturalness because humans typically use intelligibility and naturalness to define speech quality and since all systems were highly intelligible, the differences in naturalness were more noticeable. Another plausible reason for this is that being asked to rate naturalness biased the listeners towards scoring the cognitive load with respect to naturalness.

In line with previous work, no correlation between the self-reported measures and the pDCTs were found for either experiment. (Exp. 1: corr =0.1 and Exp 2. corr=0.01 with $p \geq 0.05$).

3.6 Summary

The results in this chapter demonstrates the difficulty in using a dual-task paradigm to detect differences in listening effort between highly intelligible speech synthesizers in quiet listening conditions. The amount of mental resources required to process both tasks concurrently appears to be insufficient to accurately measure listening effort. In the experiments carried out in this chapter, it is likely that participants performed well in both tasks without compromising performance on either task and therefore for no notable deterioration in the secondary task was observed.

Previous cognitive load studies have consistently reported that listeners find synthetic speech more cognitively demanding to listen to compared to human. Yet in both experiments presented in this chapter, differences were either not detected at all or the opposite result was observed. This is likely due to previous studies evaluating rule-based TTS systems which were of poorer quality (in both intelligibility and naturalness) than the systems evaluated in this work.

Listening to clean and highly intelligible synthetic speech in quiet is perhaps not cognitively demanding enough. Therefore, finding a secondary task that is cognitively demanding enough still remains a challenge. Finding a suitable secondary task can become exhaustive and this is probably the reason most studies make the primary task more challenging when evaluating human speech.

This is typically done by performing the listening task under noisy conditions [Desjardins and Doherty, 2013, Fraser et al., 2010] and therefore the same can be done for evaluating the listening effort of synthetic speech. Furthermore, to verify that the primary task was performed equally well across both tasks, researchers who wish to implement the same paradigm should consider calculating the recall accuracy formally across both experiments. In this work, recall accuracy was only performed informally. Evaluating synthetic speech in noisy conditions is important, however, the aim of our investigations at this stage was to identify a potential methodology that would be sensitive enough to detect differences in cognitive load between various speech synthesizers in quiet. Thus, using noise was avoided for the time being. Implementing the dual-task with a more challenging secondary task was an alternative method that could have been explored but this too could become exhaustive with respect to finding the most appropriate task and was therefore deemed impractical. Therefore at this point there is still an obvious need for a better methodology.

Chapter 4

Pupillometry

In this chapter, pupillometry is explored as a potential method for measuring cognitive load of synthetic speech processing. The aim is to answer the research question (RQ1) on whether such a method is sensitive to differences in listening effort between various speech synthesizers and human speech. In other words, investigating whether the pupil response reflects statistical differences in its behaviour that indicates that listening effort across various speech synthesizers exist. Previous work in this field (discussed in the next section), have successfully shown that the pupil response is sensitive to the quality of a speech signal when listening to it. Therefore, we predict that we will be successful in detecting cognitive load differences between the various speech synthesizers compared in this work. We expect to see that high quality synthetic speech produced by the Hybrid synthesizer will demand less cognitive load than speech produced by the more traditional TTS systems such as SPSS using HMMs and Unit Selection. Thus, we expect to see that as improvements on TTS systems have been made over the years, listening effort will decrease accordingly. However, we also predict that cognitive load for TTS will still demand a higher cognitive load than that of human speech as synthetic speech produced by Hybrid speech synthesizers is not as natural and as intelligible as human speech yet. We start this chapter by describing what pupillometry is, followed by a survey of work that has applied this measurement in related speech understanding studies. We then describe our methodology and implementation of pupillometry whilst discussing the same key considerations taken by us and other investigators in the field. Our results are then presented and we conclude this chapter by summarizing the key findings.

This chapter is an expansion of Govender and King [2018b] and Govender et al. [2019b].

4.1 Introduction

The pupil is considered to be a window to cognitive processing and thus many fields of research (for example in language processing, speech production and visual perception) have begun to investigate the pupil response in relation to speech understanding [Zekveld et al., 2014]. The measurement of changes in the pupil response as a function of cognitive processing is called pupillometry. Pupil data is typically collected non-intrusively using an eye-tracker or pupilometer that reflects light into

the eye in order to gauge an estimate of the pupillary size. Pupillometry studies dating back to the 1960s by Hess and Polt [1964], reported that the pupil dilates when solving arithmetic problems and that the extent of the dilation was correlated with the difficulty of the problem. Since then, findings have consistently shown that a correlation exists between pupil dilation and the mental effort required to carry out a specific task, ie., a task-evoked pupil response (TEPR).

Kahneman and Beatty [1966] presented the first study investigating the TEPRs in relation to processing performed by the short term memory. Results showed that TEPR correlated with the amount of information that is stored in memory. This was observed during a digit recall task. For every digit stored in memory, the pupil dilates and whilst the participant recalls each digit the pupil diminishes towards the baseline. This is described as a loading and unloading effect. Furthermore, they found that when participants are asked to perform a transformation on the digits, a greater pupil response is observed than the response during the recall task. This result indicates that the pupil is sensitive to task difficulty. Another study by the same investigators, showed that perception of stimuli can be measured using TEPR. This was done by measuring the pupil response whilst participants perform a pitch discrimination task [Kahnemann and Beatty, 1967]. The results showed that the pupil substantially dilates immediately after the presentation of a tone. Their findings confirmed that TEPR is correlated with the difficulty of the discrimination.

Ahern [1978] performed the first experiment that investigated pupil responses during speech perception and comprehension. In their study, participants were expected to discriminate between words that were similar or opposite in meaning. Results showed that the pupil response is twice as large for words which are not easy target words. This suggests that the pupil response is sensitive to semantic information and that it is more effortful to process difficult words than easy words. Furthermore, Beatty [1982] examined the effects of syntactic and semantic organization of sentences on the pupil response. They compared standard sentences which are described as syntactically and semantically correct with anomalous sentences that are syntactically correct but semantically incorrect and scrambled sentences that are both semantically and syntactically incorrect. For the scrambled sentences, the pupil response is greater than the anomalous sentences whilst standard sentences evoke the smallest response of them all.

Wright and Kahneman [1971] reported that the pupil response reflects local changes at intervals of the task depending upon when processing takes place. For example, whilst participants listen to a sentence the pupil dilates and peaks at the point of retention. If a question is asked immediately after, another dilation is observed which reflects the mental effort exerted to formulate an answer. If a question is asked before the sentence is presented, the pupil rapidly increases during the portion of the sentence that is related to the answer.

Having an online measurement that shows changes in pupil response at different intervals of listening makes pupillometry an attractive measurement. Recently, pupillometry has become a popular measurement for quantifying listening effort when listening to human speech under adverse conditions [Koelewijn et al., 2012, Zekveld and Kramer, 2014, Zekveld et al., 2014]. Timing is an essential part of understanding listening effort because listening demands rapid auditory encoding of speech as well as a deliberate allocation of mental resources distributed over time. With regards to

degraded signals, Zekveld et al. [2010] showed that the pupil dilates systematically with decreasing speech intelligibility .

Joshi et al. [2013] have shown that TEPR reflects activation of the locus coeruleus norepinephrine (LC-NE) system, which plays an important role in controlling autonomic functions in our brain. The LC-NE system has been associated with several cognitive functions such as the working memory, attention, reward anticipation and decision-making.

On one hand it is clear that findings have been coherent in relating the pupil response to mental effort and task difficulty but on the other hand, other studies have found correlations between the pupil response and arousal, anxiety, alertness and attention. This raises a potential concern as one can not be entirely certain as to what exactly the pupil is indexing. Nevertheless, it appears that these factors are somewhat related. It is important to realise that the pupil response is not a monotonic index of listening effort but one that comprises a combination of factors that reflects the combined contributions of the autonomic nervous system [Zekveld et al., 2018]. For example, you would expect that a person is in a high state of alertness when carrying out a mentally demanding task. To perform the task, they would also need to direct their attention toward the active processing of information. Furthermore, the pupil could be reflecting an emotional response as high levels of anxiety can also be associated with more challenging problems. Attempting to separate these factors entirely from on another remains a challenge when examining the pupil response. Since the focus of this work is not to understand each factor separately but rather as the union of these factors under the term *listening effort*, it was imperative that special attention was paid when designing our experiments to ensure that we controlled for as many other necessary factors that could potentially influence our results.

Typically, a large evoked pupil response suggests increased listening effort. However, one should not assume that more effort (or larger pupil response) always has negative implications. For example, an increased pupil response could also mean increased engagement or willingness to perform a given task. In speech communication, engagement is a process that can be viewed as a favourable one that is productive and satisfying [Winn et al., 2018].

Furthermore, Winn et al. [2018] pointed out that the pupil response is non-linear and a small pupil dilation can be evoked not only when performing an easy task but also for a hard task when effort is voluntarily withdrawn by the listener. For example, in the situation where a person is overly fatigued, there is an increased likelihood that effort will be reduced because of less engagement and as a consequence a reduced pupil size is evoked [Wang et al., 2018]. Pupil dilation is therefore said to be an index of a person's willingness to exert more effort because it is worth the exercise of greater mental resources to achieve a goal [Winn et al., 2018]. Therefore when one analyses pupil data, it is crucial to be aware of whether changes in pupil dilation are truly indicative of changes in task-related effort or unintended participant fatigue or disengagement.

To our knowledge, ours is the first attempt to measure the listening effort of synthetic speech using pupillometry. The work presented is this chapter is a starting point for determining whether pupillometry is a viable measure for this purpose.

4.2 Methodology and Implementation

Key considerations Winn et al. [2018] provided an overview of the best practices taken to set up a pupillometry experiment for the purpose of quantifying listening effort. These guidelines were taken into consideration for the implementation of our pupillometry experiments in this thesis. It is important to ensure that the audio stimuli are engaging enough otherwise participants may become disengaged. At the same time, the audio stimuli should not be too difficult otherwise listeners may abandon the task altogether. Audio stimuli selection is therefore a crucial step towards obtaining accurate results. The motivation and willingness of the participant to carry out the task also influences the pupil response and therefore if a participant is bored it is likely that their mind will wander and this will influence the validity of the results. Emotional stimuli that evoke pleasure, disgust, or any other strong physiological response should be avoided to reduce unwanted variability in the pupil data. Verbal responses increase pupil dilation and so it is recommended by other investigators that the pupil response is given sufficient time to return to baseline (pupil size at the beginning of the trial) before any verbal response is given. Physical motion is also said to influence the pupil response and therefore participants should be instructed to sit still whilst the experiment is in progress. Changes in luminance results in a change in pupil dilation that far surpasses the change evoked during cognitive tasks. Therefore, it is critical to control the visual field and surrounding light in the room when measuring the pupil response. Fatigue is avoidable for most listeners if the experiment lasts for a duration of no more than 2 hours. Fatigued listeners will show a weakened pupillary response and therefore long tedious experiments should be avoided.

Set-up Our pupillometry set-up is illustrated in Figure 4.1. Participants were instructed to fixate on a black cross displayed in the centre of the computer monitor for the duration of the trial. This was to maintain the participant's gaze whilst they listened to the audio stimulus through headphones. The participants 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 and word-error rate (WER). In addition, subjective ratings were taken at the end of each block (which comprises 20 trials). Participants were asked to rate the overall naturalness of the the audio samples (in the block they just listened to) and the difficulty they experienced in listening to them, each on 5-point scales labelled 1 - unnatural to 5 - natural, 1 - very easy to 5 - very difficult respectively. Participants were allowed to take a break between blocks if desired. The experiment lasted approximately 30-45 minutes per participant.

Pupil size data collection Pupil data was collected using an SR-eyelink 1000 plus eye-tracker. The eye-tracker was used in remote desktop mode to allow the participant to move their head freely and therefore discomfort experienced when using a head mount was avoided. The eye-tracker was positioned in front of the participant beneath a computer monitor. Only one eye was tracked. Pupil measurements were collected at 500Hz. All experiments were administered in an eye-tracking lab where lighting and sound was controlled. Each participant was calibrated with the eye-tracker at

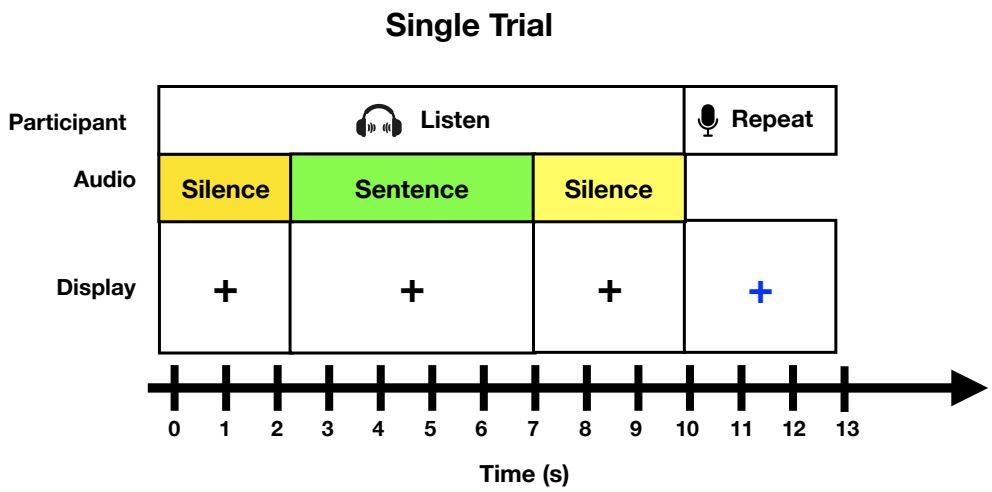


Figure 4.1: Illustration of pupillometry set-up of a single trial

the beginning of the experiment. This involved 9 black dots appearing at random positions on the computer monitor and the participant was asked to follow them with their eyes.

Audio samples Audio stimuli presented to the participants were sentences generated by the same four speech synthesizers used in the previous chapter for the dual-task paradigm in Table 3.1. For pupillometry experiments, a baseline pupil size is required for analysis. Therefore, each audio stimulus requires a buffer of at least two seconds immediately before the sentence in the audio is spoken. Depending on the experimental conditions, this buffer can be 2 seconds of silence if administered in quiet or 2 seconds of noise if administered in the presence of noise. The pupil response is considered to be slow and therefore takes some time to return to the baseline after listening has taken place. Therefore, a three seconds buffer is typically placed at the end of the audio to give the pupil sufficient time to return to the baseline. As shown in Figure 4.1, we used a 2 s and 3 s buffer at the start and end of each audio sample respectively. To ensure a fair comparison of all stimuli, loudness normalisation was subsequently applied to all audio samples using the standard root-mean-square algorithm.

Presentation Audio stimuli were presented in blocks according to the number of speech conditions being evaluated plus a practice block at the start of the experiment. Five blocks were used, one for each of the five speech conditions evaluated. Blocks were arranged using a 5 x 5 Latin square design to ensure all listeners, systems and sentences were equally balanced. The practice block comprised of 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 blocks. Each of the subsequent blocks had 20 trials. All sentences within each block were randomized except the first 5 sentences which were kept fixed across participants as they are discarded during the analysis (explained later).

Participants Participants were recruited from university students and staff, ranging in age from

19 to 37 years. All participants were native English speakers with no self-reported hearing problems.

Pupil data processing Before analysis is carried out, some pre-processing is performed on the raw pupil data. The first step is deblinking. Blinks are natural and unavoidable in such experiments. An important pre-processing step is deblinking which involves identifying and removing blinks that are typically less than 200 ms in duration. These blinks can be interpolated without interfering with the overall pupil response shape. Klingner et al. [2011] conducted an analysis concerning the effect of blinks on the pupil response and reported that difference in statistical results did not change when blink correction algorithms were applied. However, it was advised that when performing interpolation, samples from 50 ms before the blink and at least 150 ms after the blink should be taken into consideration in order to avoid task-uncorrelated high-frequency changes in the pupil response. The same procedure for deblinking was followed in our work.

Trial Exclusions The first 5 trials from each block were excluded. Investigators in Wendt et al. [2016] advised that baseline levels are substantially higher during the onset of a testing session but quickly stabilize after roughly five trials. 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. Therefore we applied the same criterion and trials with a WER greater than zero were excluded. Another criterion that was applied was to eliminate outliers. Outliers were detected using 1.5 times the interquartile ranges at 25% and 75% of data. This rule was applied by finding the mean pupil size for each trial for a given participant and if a trial mean was found to be an outlier then it was removed. Furthermore, trials that contain excessive blinks or blinks that have a duration greater than 200 ms were excluded as they cause unwanted artefacts in the trials [Klingner et al., 2011].

Baseline Correction The most common method for quantifying pupil dilation is not to report absolute pupil sizes but instead to report changes in pupil size relative to the baseline [Beatty et al., 2000]. Reporting a normalised baseline-subtracted pupil size is common. This normalises the pupil response across individuals where differences are bound to exist. This is referred to as the event-related pupil dilation (ERPD) and the percentage change from the baseline is calculated by subtracting the baseline pupil size from each pupil size sample in the trial as follows:

$$ERPD\% = \frac{sample - baseline}{baseline} \times 100 \tag{4.1}$$

The pupil size in our experiments are measured in terms of the pupil area. As mentioned earlier, a buffer is added before the onset of the spoken sentence and typically the mean pupil size during this period is used as a baseline. However, in our experiments the baseline period contained a lot of variation, possibly because the participant's mind was wandering during this period. Therefore the baseline was instead taken as the pupil size at the point immediately before the onset of the sentence.

Post-processing A 5-point moving average filter was applied to smooth the data in each trial before proceeding to the analysis.

4.3 Analysis

4.3.1 Peak Picking Analysis

Peak picking analysis is a traditional data analysis in work analysing pupillometry data [Piquado et al., 2010, Zekveld and Kramer, 2014, Schwalm et al., 2008]. The peak pupil dilation and peak latency for each participant is computed and techniques like t-tests or analysis of variance (ANOVA) with repeated measures are used to compare experimental conditions for statistical differences. Peak pupil dilation is defined as the highest value in the trial. The peak for each participant is identified either visually or automatically by finding the maximum value within a specified interval and the mean peak pupil dilation is calculated by averaging the pupil dilation over the same specified interval. In our work we computed the peak automatically in the interval the sentence was spoken between 2 to 5 seconds into the speech. Once a peak (y-value) is identified, the corresponding x-value is the time at which the peak took place and this is referred to as the peak latency.

ANOVA with repeated measures is a technique that is used to determine whether there are any statistically significant differences between the means of one or more independent groups that are based on repeated observations collected by different individuals. In our experiments, the independent groups are our speech conditions and the same participant listens to each of the five speech conditions. Therefore, the same people are being measured more than once on the same dependent variable which makes it a repeated measures design.

The repeated measures ANOVA tests for whether there are any differences between related participants means. The null hypothesis (H0) states that the means are equal:

$$H0 : \mu 1 = \mu 2 = \mu 3 = \ldots = \mu k \tag{4.2}$$

where μ = mean and k = number of speech conditions. The alternative hypothesis (HA) states that the related participants means are not equal:

$$HA : \textit{at least two means are significantly different} \tag{4.3}$$

A p-value of less than 0.05 was used to reject the null hypothesis. If the null hypothesis was observed, a post-hoc Tukey test with Bonferroni correction was used to determine which speech condition pairs are significantly different.

4.3.2 Growth Curve Analysis

Growth Curve Analysis (GCA) is a statistical technique that is typically used for analyzing time series data [Mirman, 2017]. In the previous peak picking analysis approach, the data is time-binned

and therefore its analysis is constrained to only those data points. GCA is considered to be a better approach as it analyzes the entire time course of data. In doing so, GCA captures changes in the shape and timing of the pupil response over time. The overall time course of the data is quantified by fitting orthogonal polynomial time terms to the data. Orthogonal polynomials are transformations of polynomials that result in each polynomial time term (e.g., linear (time1), quadratic (time2), cubic(time3), etc.) becoming independent from one another [Mirman, 2017]. Using orthogonal polynomials makes each time term easier to interpret. In our analysis, we fitted orthogonal polynomials up to an order of three because these time terms are well understood in the literature with respect to pupil data. The intercept refers to the overall mean pupil response where larger values indicate greater area under the curve (the fitted polynomial of the pupil size over time). The linear term refers to the slope of the pupil response where larger values indicate a steeper rising slope. The quadratic term refers to the shape of the primary curve inflection point where a high value indicates a sharper peak. The cubic term generally reflects the extent to which there is a secondary inflection point in the pupil curvature and a higher value indicates a steeper falling slope.

Using multi-level regression, we model the fixed effects of speech condition on all time terms to determine how speech condition alters the shape of the pupil response. To determine whether the results generalize to the population and sentence material, participants and item (sentence stimulus) are used as random effects on all time terms. The random effects of our model were chosen by following the recommendations of Mirman [2017]. We start with all possible random effects including trial index, block index, group index (permutation of block arrangement), participant index, sentence index, and duration. We then systematically remove random effects that either do not contribute significantly to model fit based on likelihood ratio tests, and/or offer little or no theoretical importance for interpreting the fixed effects. This process was an iterative process until a suitable model converged. The final converged model included the highest order interaction term of interest across both subjects and items. All experiments in this chapter all reduced to the same converged model structure which was as follows:

$$\text{ERPD} \sim (\text{time1} + \text{time2} + \text{time3}) * \text{CONDITION} + (\text{time1} + \text{time2} + \text{time3} \mid \text{SUBJECT}) + (\text{time1} + \text{time2} + \text{time3} \mid \text{ITEM})$$

Later in the results section we report parameter estimates using maximum likelihood estimation. These estimates are the coefficients of the closest fitted model to the raw data for each time term for each of our experimental manipulations. We interpret the results by comparing these parameter estimates in order to assess the impact that each speech condition has on the pupil response. In our work, the reference level was changed by cycling through each of the 5 speech conditions so that differences between all pairs of conditions could be identified. Statistical significance 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. The interpretations with respect to parameter estimates reported in Kuchinsky et al. [2016] are summarized in Table 4.1.

Table 4.1: Summary of interpretation of each time term in GCA. LE: Listening Effort

Term	**Characteristic**	**Interpretation**
Intercept	Area under the curve	High value → Greater mean → High LE
Linear (time1)	Rising slope	High value → Steeper slope → → High LE
Quadratic (time2)	Shape of peak	High value → more peaked pupil dilation → High LE
Cubic (time3)	Falling slope	High value → Steeper slope → High LE

High listening effort is associated with a high mean, steep rising and falling slope and a more peaked pupil response, whilst the opposite properties are true for low listening effort. These interpretations are important for understanding the results presented later in this chapter.

4.4 Experiments

This chapter comprises of three experiments that aim to determine the sensitivity of the pupil response when listening to both human and synthetic speech with the intent of answering the first research question, "Can a suitable paradigm be developed that is capable of detecting differences in cognitive load between various TTS systems."Experiment 1 investigates how the pupil response is influenced when listening to semantically unpredictable sentences (SUS). Experiment 2 compares the effects on the pupil response when listening to semantically meaningful sentences (SMS) versus listening to SUS. Experiment 3 investigates the influence on the pupil response when listening to speech in the presence of speech-shaped noise. In each experiment, we discuss results in relation to measuring the listening effort of natural and synthetic speech.

4.4.1 Experiment 1: Semantically Unpredictable Sentences

In this experiment, participants' pupil responses were collected whilst listening to audio samples containing spoken sentences that were semantically unpredictable. For example, "The moon soared through the hour glass." SUS sentences are used in traditional text-to-speech evaluations to gauge how intelligible the speech is. Sticking with this traditional method, we decided to begin our investigations with the use of SUS sentences. The SUS sentences used in both experiments can be found in Appendix A. Furthermore, listening to natural speech is considered to be effortless in ideal (quiet) conditions [McGarrigle et al., 2014]. This warranted concern that the pupil may not give a measurably large response due to insufficient load placed on the cognitive processing system of the listener. In Beatty [1982], SUS were reported to evoke a greater pupil response than simple sentences. Thus, the use of SUS sentences in this experiment was further motivated.

To our knowledge, ours is the first attempt to measure the listening effort of synthetic speech using pupillometry. Since this experiment is the starting point for determining whether pupillometry is a viable measure for the purpose of measuring listening effort, we felt that it was important to cross-validate our findings across two separate datasets. This experiment is divided into two sub-

experiments: Exp. 1A and Exp. 1B. The first evaluates the same speech conditions (selected from the 2011 Blizzard Challenge) that were evaluated using the dual-task paradigm in Chapter 3 and summarized in Table 3.1. The second evaluates four additional speech synthesizers taken from the 2010 Blizzard Challenge [King and Karaiskos, 2010] and the corresponding human speech used to build them. The 2010 Blizzard Challenge dataset comprises 5 hours of speech spoken by a British RP accented male voice talent. Synthesizers evaluated in this experiment were selected based on their naturalness and intelligibility scores such that an array of speech synthesis quality could be compared whilst still obtaining high intelligibility scores. The chosen synthesizers are summarized in Table 4.2 together with the naturalness (MOS) and intelligibility (WER) scores reported in King and Karaiskos [2010, 2011]. The synthesizers were deliberately selected according to their performance in the Blizzard Challenge. Thus our predictions for this experiment is that we would observe that natural speech demands the lowest cognitive load ie., have the lowest pupil mean, peak pupil dilation and low slope gradients compared to the speech synthesizers. As was observed in performance in the Blizzard challenge, we expected Hybrid speech to demand the least CL amongst the speech synthesizers followed by Unit Selection and HMM and the Low-Quality HMM demanding the most CL.

Table 4.2: Summary of selected speech synthesis systems, from the Blizzard Challenge 2010 and Blizzard 2011 with their naturalness (Median Score – higher is better) and intelligibility (Word Error Rate, WER – lower is better)

Speech Condition	**Blizzard Challenge 2010**		**Blizzard Challenge 2011**	
	MOS	**WER%**	**MOS**	**WER%**
Natural (Human)	5	12	5	16
Hybrid	4	19	3	20
Unit Selection	3	25	3	22
HMM	2	17	3	20
Low Quality HMM	2	18	3	26

Pre-processing

Pupil data can be messy and unpredictable and therefore requires pre-processing prior to the analysis being carried out. This is a vital step to ensure that the data we analyse is as clean as possible for the purpose of obtaining accurate findings. During pre-processing, participants can be excluded if they don't meet the relevant criteria (refer to Section 4.2 for criteria). Furthermore, trials are excluded if the pupil data of a given trial does not meet criteria. Recall accuracy is an important indicator of how difficult the listening task is in general. If high recall accuracy is obtained, this tells us that the participants were able to perform the task reasonably well and thus should not lead to fatigue. Table 4.3 summarizes the details pertaining to the pre-processing carried out for both sub-experiments.

Table 4.3: Experiment details of each sub-experiment, including the total number of participants that were included in the analysis with the total number of recruited participants in brackets, the total number of trials remaining after the trial exclusion criteria were applied with its respective percentage shown in brackets and the mean recall accuracy percentage

Experiment	**Participants**	**No. of trials (%)**	**Mean Recall Accuracy %**
1A	13(15)	808 (72)	97
1B	13(15)	768 (68)	96

Results: Intelligibility

In both of our sub-experiments, the mean recall accuracy was above 95%. This result is much better than the results achieved in the original Blizzard challenges from which the various systems were selected from. A possible reason for this is that in our experiments the errors in recall were captured manually first and then converted into a digital format that the analysis tool then used to calculated the WERs. However, in the Blizzard Challenge, the transcriptions were captured via a web-browser in which the listener was required to enter the transcriptions themselves. This can introduce spelling errors which would then be counted towards an incorrect score, resulting in increased WERs. The results in Table 4.3 confirms that all speech conditions evaluated were highly intelligible. The WERs in our experiments for each speech condition evaluated is presented in Table 4.4. In Exp. 1A, Low-Quality HMM has the highest WER and is significantly different to all other speech conditions. In Exp. 1B, Unit Selection has the highest WER and is significantly[1] different to all other speech conditions except Low-Quality HMM. Natural speech has the lowest WER in both sub-experiments but was only significantly different to the speech conditions that have the highest WERs. The conditions that have the highest and lowest WERs in our experiment correspond with those reported in the respective Blizzard Challenges in Table 4.2. There is a possibility that the lack of significant results can be attributed to the power of the sample size (N=13).

Table 4.4: WER of speech conditions in Exp. 1A and Exp. 1B

Speech Condition	**WER %**	
	Exp. 1A	**Exp. 1B**
Natural (Human)	2	1
Hybrid	3	3
Unit Selection	2	9
HMM	2	3
Low Quality HMM	5	6

Results: Self-reported measures

The medians of the self-reported measures are presented in Table 4.5. In Exp. 1A, listeners found Natural speech to be the most natural sounding and the easiest to listen to. Low-Quality HMM

[1]Please note: Statistical significance was calculated on the basis of the null hypothesis being rejected if p ¡ 0.05. All statistical results can be found in Appendix B.

Table 4.5: Self-reported measures, Naturalness Median Score – higher is better) and Cognitive Load (CL) – lower is better)

Speech Condition	**Exp. 1A**		**Exp. 1B**	
	MOS	**CL**	**MOS**	**CL**
Natural (Human)	4	2	4	1
Hybrid	3	2	3	2
Unit Selection	2	3	2	4
HMM	2	3	2	3
Low Quality HMM	1	4	2	3

was rated as the least natural and most difficult to listen to. Hybrid was the most natural sounding and easiest to listen to synthetic speech. In all cases, these results were statistically different to all other speech conditions except Low-Quality and Unit Selection (US) which were equivalent in cognitive load. Apart from the Low-Quality HMM being equivalent to Unit Selection, the general expected trend from high to low was observed.

In Exp. 1B, Natural speech was rated the most natural sounding and easiest to listen to. Hybrid was the only condition equivalent to Natural in both naturalness and cognitive load, meaning no significant difference was found between Natural and Hybrid speech. Unit Selection, HMM and Low-Quality HMM were equally unnatural and equally difficult to listen to.

A significant negative correlation between naturalness and cognitive load was found (Exp. 1A: corr=-0.53 and Exp. 1B: corr=-0.60). This finding indicates that participants found the more natural-sounding speech conditions to be easier to listen to and vice versa.

Results: Peak picking ANOVA analysis

Table 4.6: Peak picking ANOVA results for mean pupil dilation, peak pupil dilation and peak latency in Exp. 1A and Exp. 1B

Experiment	$df1$	$df2$	**Mean**		**Peak**		**Latency**	
			F	p	F	p	F	p
Exp. 1A	4	48	0.83	0.51	1.21	0.32	2.59	**0.05**
Exp. 1B	4	48	0.34	0.85	0.51	0.73	0.23	0.92

The peak picking ANOVA analysis (refer to Section 4.3.1) results are shown in Table 4.6. We expected the mean and peak pupil dilation to have a significant effect on the pupil response and for peak latency to have no effect. In Exp. 1A, results indicated that speech condition has no significant effect on the mean pupil response or peak pupil dilation and the same was true in Exp. 1B. However, speech condition did have a significant effect on peak latency in Exp. 1A. The pupil response when listening to Natural speech was significantly delayed in comparison to evoked pupil response when listening to the Low-Quality HMM system. Our hypothesis for this main effect was that the speaker in Exp. 1A spoke in a manner that was over articulated and therefore the sentences for natural speech were generally longer than synthetic speech. As a consequence, this could have resulted in a delayed peak. To test this we computed the mean duration of the audio samples

for Natural and Low-Quality HMM and measured the correlation between the mean durations and peak latencies. However, no significant correlation was found. Reasons for this delay are yet to be understood.

Results: Growth Curve Analysis

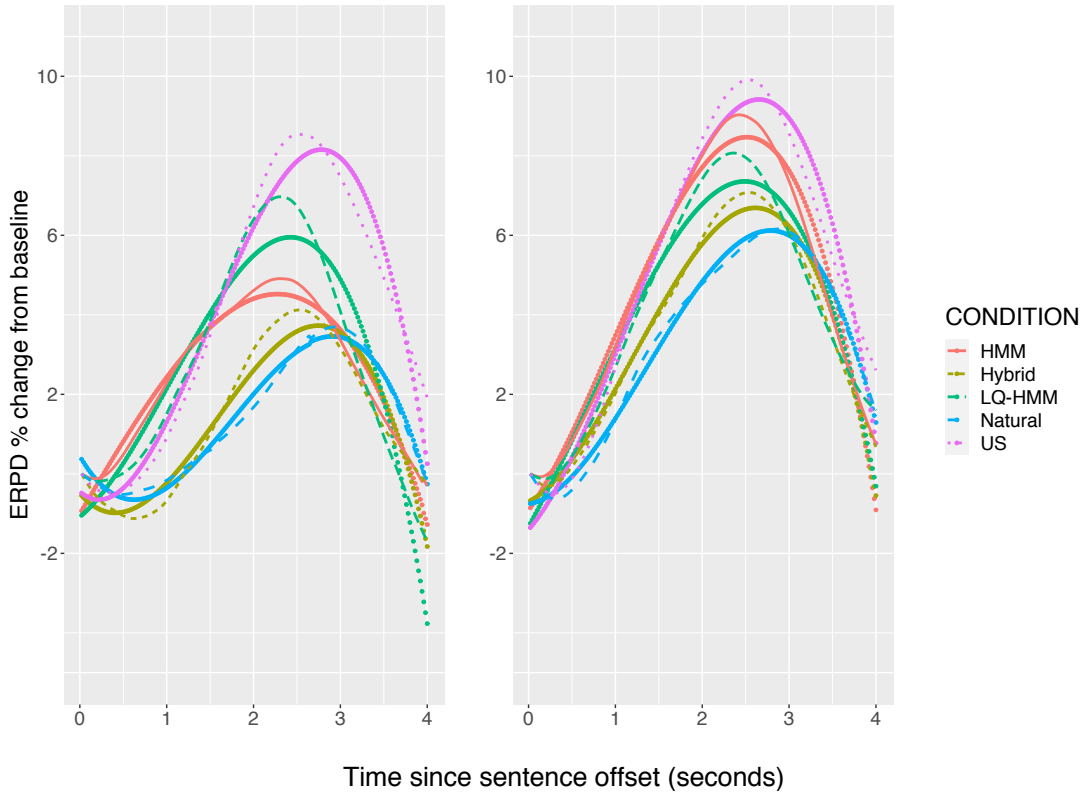


Figure 4.2: Time series line graph of the raw pupil response (dotted) and cubic model fit (solid line) averaged across all participants, Exp. 1A shown on left and Exp.1B shown on right

The ERPD % change from the baseline for the raw pupil responses and cubic model fits are shown in 4.2. Following the procedure described in Section 4.3.2 using likelihood ratio testing, the model comparisons showed a significant[2] effect of speech condition on all time terms (linear, quadratic and cubic). All results described in this section refer to differences with respect to the parameter estimates using maximum likelihood estimation and are shown in Table 4.13. Findings are described with reference to the interpretations that are described in Table 4.1. Our hypothesis for this experiment was that the intercept, linear and quadratic terms would all have low parameter estimates for Natural speech which can interpreted as Natural speech demanding the lowest CL. In addition, we expected to see increasing CL as the quality of the TTS systems deteriorated from Hybrid to Unit Selection to HMM to Low-Quality HMM. Therefore, expecting to see increasing parameter estimates as we go from the highest quality speech synthesizer to the lowest quality speech synthesizer.

[2]Please note: All statistical results can be found in Appendix B.

Table 4.7: GCA maximum likelihood parameter estimates of each time term for each evaluated speech condition in Exp. 1A

Condition	Intercept	Linear	Quadratic	Cubic
Natural	1.32	13.99	-8.34	-11.39
Hybrid	1.34	12.28	-17.06	-12.09
HMM	2.83	6.53	-23.46	-3.77
Low-Quality	3.00	7.60	-33.67	-11.88
Unit Selection	3.39	25.46	-23.11	-16.41

Table 4.8: GCA maximum likelihood parameter estimates of each time term for each evaluated speech condition in Exp. 1B

Condition	Intercept	Linear	Quadratic	Cubic
Natural	3.09	24.06	-17.22	-10.49
Hybrid	3.70	19.81	-27.22	-12.61
HMM	4.10	12.80	-34.21	-11.18
Low-Quality	3.89	15.08	-31.05	-8.75
Unit Selection	4.43	23.33	-34.47	-13.76

Intercept In Exp. 1A, Natural and Hybrid have the lowest means and are equivalent. Unit Selection has the highest mean and is significantly[3] different from all other speech conditions. HMM and Low-Quality HMM are equivalent. These results suggest that Unit Selection demands the most listening effort. This is an interesting result because Unit Selection was not reported as the most unnatural or most difficult to listen to speech condition in the self-reported measures, neither did it have the highest WER. Natural and Hybrid have the lowest means and therefore demand the least amount of listening effort in line with the listeners' perceived cognitive load and our predictions.

In Exp. 1B, we observe similar results. Natural speech has the lowest mean and Unit Selection has the highest. Hybrid has the second lowest. All speech conditions were statistically different to one another in Exp. 1B whilst in Exp. 1A, conditions group together forming rankings of the lowest, highest and in-between. Since the Low-Quality HMM condition did not result in the highest mean in Exp. 1A, we are led to believe that listening effort is not influenced by intelligibility alone but rather a combination of various factors. For Unit Selection specifically lack of naturalness could contribute to increased cognitive load due to artefacts introduced by joining speech units from various contexts together where they may not necessarily belong and digital processing performed to smooth the signal also contributes to deterioration in naturalness. Another interesting observation is that the parameter estimates for the intercept are higher for all conditions in Exp. 1B than in Exp. 1A. This finding suggests that participants exerted greater mental effort when listening to natural and synthetic speech that is spoken in Exp. 1B compared to Exp. 1A. This could be a

[3]Please note: All statistical results can be found in Appendix B.

result of numerous differences between these two datasets such as gender, accent, speaker characteristics and/or sentence material which directly influences intelligibility and thus requires further investigations beyond the scope of this thesis.

Linear term In both Exp.1A and Exp. 1B, HMM and Low-Quality HMM are equivalent and have the flattest slopes whilst in both sub-experiments Unit Selection has the steepest slope. With respect to Unit Selection, this finding aligns with those reported in Kuchinsky et al. [2013] that more difficult listening conditions have steeper slopes. In other words, Unit Selection demands more listening effort than any other condition. On the contrary, HMM and Low-Quality HMM in both sub-experiments have the flattest slopes but are not the conditions reported by listeners' as the easiest to listen to. In addition, apart from Unit Selection, Natural and Hybrid speech in both sub-experiments have the steepest slope. This leads us to believe that the slope may be indexing some other cognitive resource that may not be associated with listening difficulty for these conditions as it is unlikely for the HMM systems to be easier to listen to than human speech which is the upper bound of this experiment and the quality we are aiming to achieve. This is also confirmed by the self-reported scores where human speech is rated the easiest to listen to compared to the HMM systems.

Quadratic term In both sub-experiments, Natural speech has the flattest peak and is significantly[4] different to all other speech conditions. Therefore, Natural speech demands the least amount of listening effort. Increasing in sharpness, the Hybrid condition followed Natural and was also significantly different to all other conditions. In Exp. 1A, Unit Selection and HMM have equivalent peak shapes and Low-Quality HMM has the sharpest peak. Therefore, according to the peak shape alone, Low-Quality HMM demands the most listening effort. In Exp. 1B, Unit Selection has the sharpest peak. Therefore, Unit Selection demands the most listening effort in Exp. 1B. The shape of the peak appears to correspond more closely with the self-reported cognitive load scores reported by listeners' in Table 3.5 as well intelligibility. Therefore, sharper peaks are associated with high listening effort and poor intelligibility whilst flat peaks are associated with low listening effort and high intelligibility. Since a strong correlation between self-reported cognitive load and naturalness measures was found, perhaps this time term is also indexing listening effort in relation to how natural the speech sounds. In other words, natural-sounding speech evokes a flatter peak compared to poor quality speech that evokes a sharp peak.

Cubic term Unit Selection has the steepest falling slope in both Exp. 1A and 1B which is associated with a high listening effort as it indicates a rapid decline in the pupil response when returning to the baseline. In Exp. 1A, the flattest slope was HMM and was significantly different to all other conditions. In Exp. 1B, Low-Quality HMM has the flattest slope. These results contradict with the results we expect, as both HMM conditions are expected to demand more listening effort than any other condition (except Unit Selection) according to other terms. We observed similar results in the linear term. Therefore, once again we are lead to believe that both the rising and

[4]Please note: All statistical results can be found in Appendix B.

falling slopes appear to be indexing a cognitive resource that may not be associated with listening difficulty for all conditions except Unit Selection.

Summary

In this experiment we set out to determine the practicality of using pupillometry to measure the listening effort of synthetic speech. More specifically, we chose speech synthesizers of varying speech quality such that we could understand how speech quality of synthetic speech influences the pupil response. Measuring both intelligibility and naturalness in this experiment was important to ensure that scores in our experiment corresponded with those of the Blizzard Challenge, as the original Blizzard Challenge scores formed the basis of our synthesizer selection process. Results for both sub-experiments and their respective Blizzard Challenges corresponded for the speech synthesizers that had the lowest and highest WERs, ie., Natural speech in both sub-experiments and Low-Quality HMM in Exp. 1A and Unit Selection in Exp. 1B. Similarly, participants reported Natural as being the most natural sounding and easiest to listen to and Low-Quality and Unit Selection as the most unnatural and difficult to listen in Exp. 1A and Exp.1B respectively. Hybrid speech was specifically selected as the most natural speech synthesizer and participants in both sub-experiments agreed. Furthermore, a significant negative correlation was found between the self-reported naturalness and cognitive load scores. This result isn't surprising. One would expect that when listening under ideal conditions like in quiet, listeners will naturally form judgement on how natural the speech sounds which is likely to influence their perception with regards to ease of listening.

Based on the subjective results alone, we would expect Natural speech to demand the least cognitive resources in both sub-experiments and Low-Quality HMM and Unit Selection to demand the most cognitive resources in Exp. 1A and 1B respectively. High listening effort as understood in the literature is associated with a high mean, steep rising slope, sharp peak and steep falling slope. Natural speech opposes all of these properties in both sub-experiments with the exception of having a steep rising and falling slope. Results in the linear and cubic terms suggest that in both these sub-experiments, a steep rising slope is indexing some other resource other than listening difficulty. The fact that we observe the same trend across both datasets increases the validity of this finding. More so, it is unlikely that Natural speech demands more listening effort than Low-Quality HMM (that has the flattest slope) especially when all other time terms suggest the opposite. Therefore, we conclude that Natural speech demands the lowest listening effort in both sub-experiments. Hybrid demands the lowest listening effort from all the speech synthesizers which was expected. Unit Selection demands the highest listening effort which is evident in all time terms. Given that we believe that the linear and cubic terms are indexing something else it is interesting that Unit Selection was the only system that consistently aligned with the expected properties of high listening effort. This was unexpected for Exp. 1A where Unit Selection demands the highest listening effort despite it being a synthesizer that does not sound the most unnatural and is highly intelligible. This tells us that the listening effort measurement is not influenced by intelligibility alone. It also implies that Unit Selection is processed differently to all other synthesizers. This finding makes sense as Unit Selection uses natural speech units to generate speech compared with

all other systems. The HMM systems showed evidence of high listening effort but differed to Unit Selection in that it evoked flat rising and falling slopes.

Overall, results show that the pupil response is sensitive to changes in speech quality. Synthetic speech imposes greater listening effort in both sub-experiments compared to natural speech. Differences between the speech synthesizers were detected. Hybrid demands the least effort whilst Unit Selection demands the most. The peak shape appears to correlate well with naturalness perception. Therefore, listening effort isn't specifically influenced only by intelligibility alone but rather a combination of intelligibility, naturalness and other properties of the speech signal that traditional listening tests do not reveal. Therefore, this experiment confirms that pupillometry is a viable method for measuring the listening effort of synthetic speech and was shown to be sensitive enough to detect differences between the selected speech synthesizers. However, the ecological validity of the results remains a concern.

4.4.2 Experiment 2: Semantically meaningful sentences

In this experiment, participants' pupil sizes were collected whilst listening to audio samples containing spoken sentences that have semantic meaning. For example, "To an extent, the council too is worn out." Our concern was that using SUS sentences is not ecologically valid. If an experiment is unable to simulate the real-world listening conditions then are they really a true reflection of how effortful synthetic speech is. SUS sentences are *unfamiliar* and therefore could potentially tap into mental resources that are different to the mental resources one may utilize when listening to common and familiar sentences. Listening to human speech in quiet is considered to be effortless and perhaps this is why investigators found it necessary to increase the load when evaluating the listening effort of human speech. Synthetic speech is however considered more challenging to listen to than natural speech and thus it may be possible to detect differences under quiet experimental conditions. The aim of this experiment was to compare the effect on the pupil response when listening to semantically meaningful sentences (SMS) of natural and synthetic speech (Exp. 2) compared to SUS (Exp. 1A discussed in previous section). Only the speech conditions from the 2011 Blizzard Challenge is used as it is the most recent of the two datasets evaluated. In this experiment we predict that when listening to SMS, differences between the higher quality speech synthesizers will not be detected but may be detected between the high and low quality speech synthesizers. For example, differences between Hybrid and Low-Quality HMM.

Pre-processing

Details pertaining to the pre-processing carried out for Exp. 2 (SMS) are summarized in Table 4.9. For ease of comparison we included details for Exp. 1A (SUS).

Table 4.9: Analysis details of Exp. 1A (SUS) and Exp. 2 (SMS), including the total number of participants that were included in the analysis with the total number of recruited participants in brackets, the total number of trials remaining after the trial exclusion criteria was applied with its respective percentage shown in brackets and the mean recall accuracy percentage

Experiment	**Participants**	**No. of trials (%)**	**Mean Recall Accuracy %**
1A (SUS)	13(15)	808 (72)	97
2 (SMS)	15(15)	869 (77)	96

Results: Intelligibility

Recall accuracy in both experiments was above 95%. Therefore changing sentence material from SUS to SMS did not have an effect on participants' ability to accurately recall sentences as the intelligibility across both experiments remained the same. Table 4.10 shows the WERs for both experiments for each speech condition. Significant[5] differences in Exp. 1A were found between Low-Quality HMM and all other speech conditions. Low-Quality HMM has the highest WER. In

[5]Please note: All statistical results can be found in Appendix B.

Exp. 2, Natural speech has the highest WER and Hybrid has the lowest. An interesting observation is that the WERs are greater in Exp. 2 for most speech conditions than in Exp. 1A. During data collection it was observed that participants had a tendency to recall sentences by repeating them in their own words as opposed to repeating the sentences verbatim despite being instructed to do so. This is an important processing difference when listening to familiar sentences versus unfamiliar sentences. When listening to sentences we understand, we naturally rearrange words into something that makes it easier for us to comprehend. For meaningless sentences, the listener is aware that it doesn't make sense and therefore is likely to repeat the words verbatim. This may have been a contributing factor in Natural speech having the highest WER in Exp. 2. We believe that the same contributing factor resulted in the WERs being higher in Exp. 2 than Exp. 1A.

Table 4.10: WER percentage of speech conditions in Exp. 1A and Exp. 2

Speech Condition	**WER %**	
	Exp.1 A (SUS)	**Exp. 2 (SMS)**
Natural (Human)	2	6
Hybrid	3	2
Unit Selection	2	4
HMM	2	5
Low Quality HMM	5	4

Results: Self-reported measures

The medians of the self-reported measures for Exp. 1A and 2 are presented in Table 4.11. In Exp. 1A, listeners found Natural speech to be the most natural sounding and the easiest to listen to and for both measures these results were significantly[6] different to all other speech conditions. The Low-Quality HMM was rated as the most unnatural and difficult to listen to and is significantly different to all other conditions except Unit Selection in cognitive load. In Exp. 2, Natural speech is rated the most natural sounding and easiest to listen to and Hybrid is the only condition found to be equivalent to Natural speech in terms of naturalness. Low-Quality HMM is the most unnatural and difficult to listen to and is significantly different to all other conditions with the exception of being equivalent to HMM in naturalness. The medians for naturalness are exactly the same in both experiments. Therefore, sentence material does not influence participants' perception in naturalness. For cognitive load, listeners found it easier to listen to SMS than SUS for Natural speech. In contrast, Hybrid which was equivalent in naturalness to Natural speech in Exp. 2 is affected by the sentence material. In other words, participants found it more difficult to listen to SMS than SUS whilst listening to Hybrid speech. The medians for Unit Selection, HMM and Low-Quality HMM remained unchanged across both experiments. The type of sentence material appears to have an effect only on the highest quality speech synthesizer. A positive effect is observed for Natural speech where cognitive load was reduced whilst a negative effect is observed for Hybrid speech where cognitive load was increased.

[6]Please note: All statistical results can be found in Appendix B.

A significant[7] negative correlation between the two self-reported measures was found in both experiments (Exp. 1A: corr=-0.53 and Exp. 2: corr=-0.68).

Table 4.11: Self-reported measures (Median Score, – higher is better) and (Cognitive Load, CL – lower is better)

Speech Condition	**Exp. 1A (SUS)**		**Exp. 2 (SMS)**	
	MOS	**CL**	**MOS**	**CL**
Natural (Human)	4	2	4	1
Hybrid	3	2	3	3
Unit Selection	2	3	2	3
HMM	2	3	2	3
Low Quality HMM	1	4	1	4

[7]Please note: All statistical results can be found in Appendix B.

Results: Analysis of Variance

Table 4.12: ANOVA results for mean pupil dilation, peak pupil dilation and peak latency in Exp. 1A and Exp. 2

Experiment	$df1$	$df2$	**Mean**		**Peak**		**Latency**	
			F	p	F	p	F	p
Exp. 1A (SUS)	4	48	0.83	0.51	1.21	0.32	2.59	**0.05**
Exp. 2 (SMS)	4	56	2.04	0.10	1.70	0.16	1.95	0.12

Similar to the previous experiment, for the ANOVA we expected to see a significant effect on the mean and peak pupil dilation between Natural and the Low-Quality HMM as well as between Hybrid and the Low-Quality HMM. The ANOVA analysis results are shown in Table 4.12. The only significant[8] result was the main effect of speech condition on the peak latency in Exp. 1A, where Natural speech was significantly delayed in comparison to the Low-Quality HMM system. It is interesting that peak is no longer significantly delayed when listening to SMS. Visually, Natural speech is delayed compared to all other conditions yet significance is not reached when using ANOVA. This leads is to believe that the delay in peak latency is due to the sentence material. In other words, processing is delayed when listening to SUS which are unfamiliar to our cognitive processing system. However, this was only observed for Natural speech.

Results: Growth Curve Analysis

Table 4.13: GCA parameter estimates of each time term and speech condition in Exp. 1A (SUS)

Condition	**Intercept**	**Linear**	**Quadratic**	**Cubic**
Natural	1.32	13.99	-8.34	-11.39
Hybrid	1.34	12.28	-17.06	-12.09
HMM	2.83	6.53	-23.46	-3.77
Low-Quality	3.00	7.60	-33.67	-11.88
Unit Selection	3.39	25.46	-23.11	-16.41

Table 4.14: GCA parameter estimates of each time term and speech condition in Exp. 2 (SMS)

Condition	**Intercept**	**Linear**	**Quadratic**	**Cubic**
Natural	2.09	25.40	-6.33	-6.82
Hybrid	5.78	35.05	-42.02	-19.46
HMM	6.55	32.18	-40.49	-8.21
Low-Quality	3.99	3.77	-38.79	-0.64
Unit Selection	5.53	27.18	-31.53	-7.76

[8]Please note: All statistical results can be found in Appendix B.

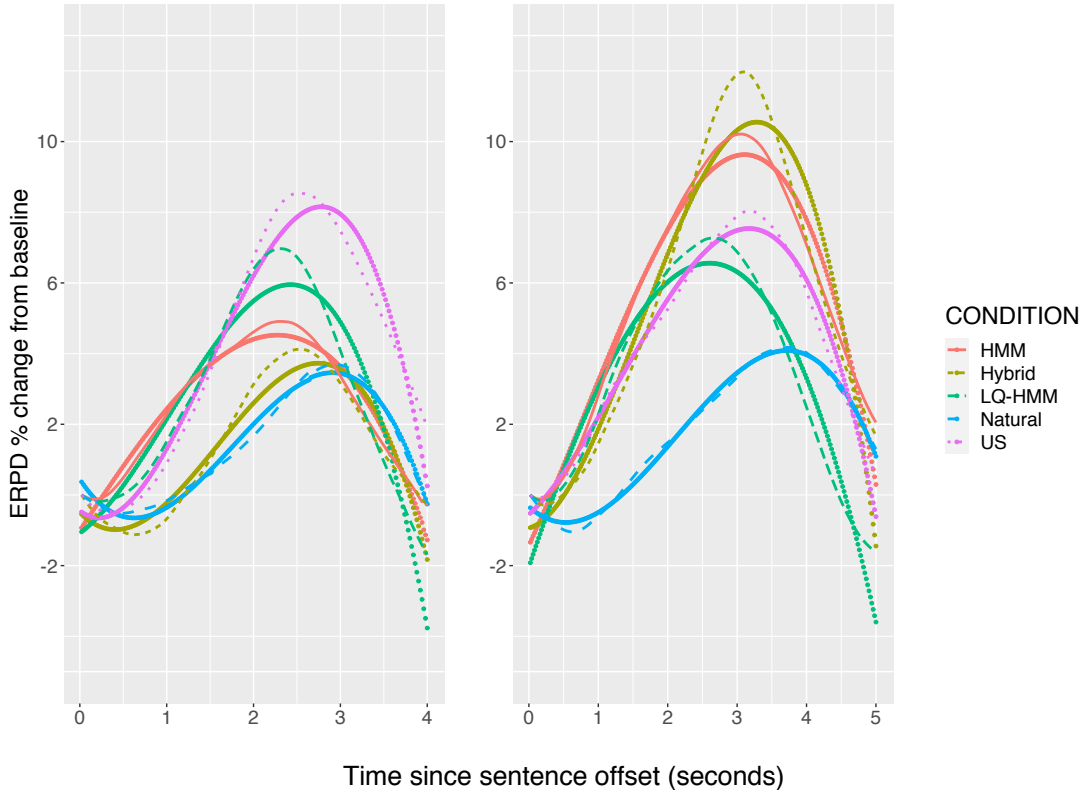


Figure 4.3: Time series line graph of the ERPD % of the raw pupil response (dotted) and cubic model fit (solid) averaged across all participants, Exp. 1A (SUS) shown on left and Exp.2 (SMS) shown on right.

For the GCA analysis, we expected to see parameter estimates to be low for Natural and Hybrid synthesizer but high for the Low-Quality HMM. We only expected to see differences between the highest and lowest quality systems as our concern was the listening to SMS in quiet may not be cognitively demanding enough to detect differences between the various speech synthesizers.

The ERPD % change from the baseline for the raw pupil responses and cubic model fits are shown in 4.3. Using the likelihood ratio testing, the model comparisons showed a significant[9] effect of speech condition on all time terms (linear, quadratic and cubic).

Intercept In Exp. 1A, Natural and Hybrid are equivalent and evoke the lowest mean pupil response. In contrast, Unit Selection evokes the greatest mean pupil response. In Exp. 2, we observe the same result for Natural. Therefore, sentence material did not influence the pupil response of natural speech. Hybrid speech, however, goes from the lowest mean pupil response in Exp. 1A to the second highest mean response in Exp. 2. Therefore, the type of sentence material appears to influence the most natural sounding speech synthesizer. With respect to the parameter estimates, it is evident in the results that all intercept values are higher in Exp. 2 than Exp. 1A. If the pupil response is indexing listening difficulty, this would suggest that SUS are easier to process than SMS which is unlikely as previous research has shown otherwise [Beatty, 1982]. This leads us to believe that in Exp. 2, the pupil response is not indexing listening difficulty in the case of listening to synthetic speech. This could explain why Low-Quality HMM, which is significantly different to all

[9]Please note: All statistical results can be found in Appendix B.

other conditions, has the lowest mean.

Linear term In Exp. 1A, HMM and Low-Quality HMM have the flattest slopes whilst Unit Selection has the steepest slopes. In Exp. 2, Low-Quality HMM has the flattest slope and Hybrid has the steepest slope. However, all estimates are steep except for Low-Quality HMM. Changing to SMS reduced the estimate for the Low-Quality HMM system whilst it increased for all other conditions. We can interpret this as Low-Quality HMM demanding the least resources. Therefore it seems intuitive that the slope reflects more the level of engagement as listening to SMS that has more meaning will likely be more engaging and therefore explains why most synthesizers and natural speech evoke steep slopes. If this is so, then one should be weary that high levels of engagement could be influenced by speech condition being highly demanding and/or highly engaging. Therefore comparison with other time terms will aid in disambiguating the differences.

Quadratic term In both experiments, Natural speech has the flattest peak which is associated with demanding the least amount of listening effort and was significantly[10] different to all other speech conditions. We observe that all parameter estimates increase in value which implies that the peaks are all sharper in Exp. 2 than Exp. 1A except for natural speech. This implies that natural speech was unaffected by the change in sentence material. Furthermore, Hybrid changes from having the second flattest peak in Exp. 1A to the sharpest peak in Exp. 2. When listening to SUS Natural and Hybrid were equivalent yet listening to SMS, the peak remains flat for Natural but becomes sharp for Hybrid. This suggests that synthetic speech is processed differently to natural speech when listening to SMS. This warrants our concern that processing unfamiliar sentences versus familiar sentences may not utilize the same mental resources for Natural speech. This motivates why previous studies used SUS sentences. Perhaps an insufficient load is placed on Natural speech when listening to SMS. By knowing that Hybrid was deliberately selected as the highest quality synthesizer in the evaluation, it seems unlikely that Hybrid would demand the most listening effort when listening to SMS if the pupil response is indexing resources associated with listening difficulty. Therefore, this once again leads us to believe that the pupil may be indexing something other than listening difficulty in the case of listening to SMS for synthetic speech.

Cubic term In Exp. 1A, HMM has the the flattest slope and Unit Selection has the steepest slope compared to all other conditions. In Exp.2, Hybrid has the steepest slope and Low-Quality HMM has the flattest slope. Once again, these properties reflect the opposite to what we expect - Listening to high quality speech demands less resources than low quality speech. In other words, Low-Quality is expected to demand the highest listening effort and Hybrid the lowest. Therefore, once again we are lead to believe that listening difficulty is not being indexed in this experiment for synthetic speech.

The effect on the pupil response when listening to SUS and SMS for each speech condition is shown separately in Figure 4.4. Based on our findings, it is evident that listening to SMS has a

[10]Please note: All statistical results can be found in Appendix B.

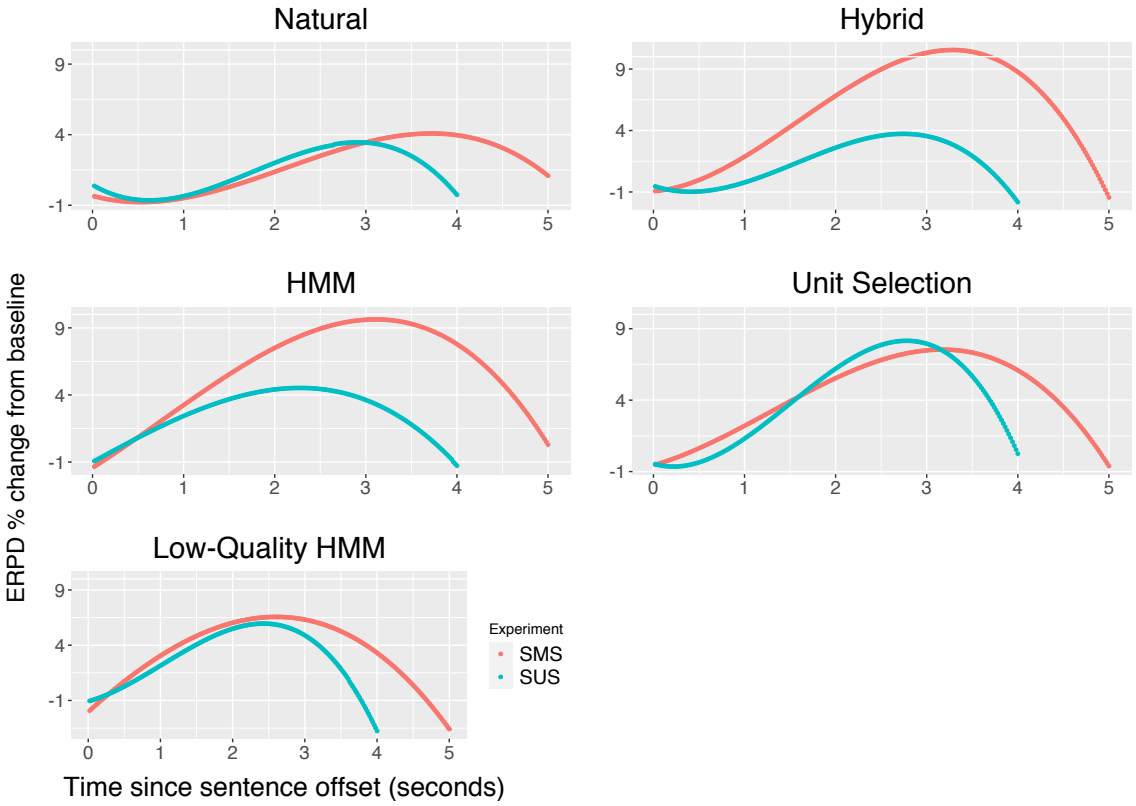


Figure 4.4: Time series line graph of cubic model fits for Exp. 1A and Exp. 2 for each speech condition individually, where ERPD % change from baseline is on the y-axis and the time in seconds is on the x-axis.

greater influence on Hybrid and HMM compared to all other speech conditions. This translates to the higher quality speech synthesizers demanding more cognitive resources when processing SMS compared with processing SUS. We expect that SUS will be more difficult to process that SMS due to the lack of semantic plausibility as was observed in [Beatty, 1982]. Thus it is more likely that different resources are being indexed by the pupil response in these experiments. Our hypothesis is that when listening to SMS produced by high quality speech synthesizers, the pupil response shows that the high quality synthesizer demands more resources than the low quality synthesizer which leads us to believe that attention resources are being index opposed to resources of the working memory. Therefore appearing to demand more attention resources related to levels of engagement. Therefore, Hybrid and HMM synthesizers demand more attention resources than any other synthesizer in this experiment and demand similar resources to process the difficulty of SUS sentences. Not much of an effect is observed between both sub-experiments for Natural speech, Unit Selection and Low-Quality HMM. The pupil response for Natural speech is unaffected by the sentence material and is low in both cases as listening to human speech is considered effortless and therefore may not belong on the same scale as synthetic speech. Low-Quality HMM - the poorest quality synthesizer - is associated with low levels of engagement when listening to SMS. In contrast, Unit Selection elicits a similar pupil response in both sub-experiments, the pupil response differs to Natural and Low-Quality HMM as its mean is comparatively higher in both sub-experiments. As previously mentioned, Unit Selection appears to be processed differently compared to any other

speech synthesizer. Unit Selection, in this experiment, appears to be associated with high levels of engagement. However, in the last experiment it demanded the most listening effort. Therefore, our hypothesis is that high levels of engagement in this case is more likely due to the listener being in a high state of alertness in order to process the speech produced by this synthesizer which is more challenging than some of the other speech synthesizers.

Summary

In this experiment we set out to determine whether sentence material influences the pupil response. Although using SUS proved viable in measuring the listening effort of synthetic speech in Experiment 1, our concern was that using SUS is not ecologically valid.

Sentence material influenced intelligibility, where all WERs were higher when listening to SMS compared to SUS. This was believed to be a result of sentence familiarity and comprehension. Participants had a tendency to rearrange words or substitute words with synonyms when listening to SMS which resulted in increased WERs. Apart from this, Low-Quality HMM had the highest WER in both experiments.

In terms of self-reported measures, sentence material had no influence in the perception of speech naturalness. This was supported by the median scores remaining unchanged in both experiments. With regards to cognitive load, participants found it easier to listen to Natural speech when listening to SMS. In contrast, Hybrid speech was found to be more difficult when listening to SMS comapred to SUS. Low-Quality HMM was reported as being the most difficult to listen to for both SMS and SUS.

The mean pupil responses for all speech conditions were greater when listening to SMS compared to SUS. This implies that the cognitive load is greater when semantics are present. However, previous studies have consistently shown that processing SUS is more difficult to process than processing SMS [Beatty, 1982]. This result therefore supports the idea that different mental resources are utilized when processing familiar and meaningful sentences versus unfamiliar and meaningless sentences. Findings suggest that when listening to SUS a high mean pupil response, sharp peak and steep rising and falling slope are all properties of high listening difficulty, all of which corresponded with the Unit Selection speech synthesizer which indeed was one of the more challenging synthesizers to process. In contrast, when listening to SMS, the Hybrid synthesizer has a high mean, sharp peak and steep rising and falling slopes, implying that the highest quality speech synthesizer (lowest WER and highest naturalness score) demands high listening difficulty when processing SMS which we believe is unlikely. This finding leads us to believe that when processing SMS, it is more plausible that the pupil response is indexing something other than listening difficulty.

In Chapter 2, we mention that the dominant cognitive processes include the working memory, attention, speed of processing and linguistic knowledge. Since we relate working memory to the difficulty of processing which we ruled out this leaves us with attention, speed of processing and linguistic knowledge. If we consider speed of processing, it makes sense that SMS would be processed faster than SUS which could explain the increased pupil responses. However, that would mean that all systems would evoke a significantly higher pupil response including Natural speech which isn't the

case and therefore can also be ruled out. With similar reasoning we can rule out linguistic knowledge which thus leaves us only with attention resources, translating to increased levels of engagement. Most time terms in our growth curve analysis support this notion. However, this appears valid only in the case of synthetic speech and not natural speech. Natural speech has the lowest mean and flattest peak in both sub-experiments whilst only the slopes align with those of synthetic speech. Therefore it is questionable whether natural speech and synthetic speech are indexing the same resources across all time terms. The peak of Natural speech was also shown to be significantly delayed when processing SUS. Since the accent, gender and speaker characteristics were the same across both experiments, we can conclude that sentence material was the only factor remaining and thus responsible for this effect. Finally, we observed that Natural speech was not influenced much when changing the sentence material other than the peak latency which suggests that different mental resources are used when processing natural speech. This supports the notion that cognitive processing for natural and synthetic speech differs when processing SMS. Thus, cognitive load for natural speech and synthetic speech may not be directly comparable in this experiment. Although differences were detected between Natural speech and synthetic speech and between the various speech synthesizers, there was still uncertainty on whether Natural speech and synthetic speech are comparable in this experiment due to uncertainties of what resource is being measured and when. Therefore this method did not prove viable for our purposes. Since the key objective is to successfully detect difference in listening effort and more specifically listening difficulty, it is important to ensure that the resources of the working memory are being indexed. To ensure listening difficulty is being indexed, we hypothesise that investigating the pupil response when listening under noisy conditions should force resources of the working memory to be indexed as is well known that listening in noise is more challenging than listening in quiet especially if no speech enhancement is being applied.

4.4.3 Experiment 3: Quiet vs Noise

In Experiment 2, we observed that when listening in quiet conditions to semantically meaningful predictable sentences, the pupil response appears to be indexing something other than listening difficulty and when listening to SUS, the pupil indexes listening difficulty. However, the use of SUS may not be ecologically valid as there is reason to believe that SUS is processed differently to SMS. To ensure a sufficient load is placed on listeners such that the pupil response indexes listening difficulty is to measure listening effort in the presence of noise [Zekveld et al., 2010, 2011, Koelewijn et al., 2012]. Simantiraki et al. [2018] 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 (SNRs). The aim of this experiment was to compare the effect on the pupil response when listening to natural and synthetic speech in quiet compared to listening in noise. For comparative purposes, the same speech conditions in Experiment 2 are evaluated in this experiment. Two sub-experiments were conducted each measuring the listening effort of natural and synthetic speech in the presence of speech-shaped noise at SNRs -1dB (Exp. 3A) and -3dB (Exp. B) respectively. These SNRs were chosen specifically such that the cognitive load is increased whilst intelligibility remains close to ceiling. Cooke et al. [2013] reported that when listening to natural speech in -1dB and -3dB, intelligibility of approximately 80% and 60% is obtained respectively. In Simantiraki et al. [2018], the TTS condition at -5dB SNR was too difficult and therefore we only evaluated the influence of noise up to -3dB SNR. Our prediction for this experiment is that listening effort differences will be detected when listening in the presence of noise. More specifically, we believe that the working memory resources will be indexed and thereby reflecting listening difficulty differences between the various speech synthesizers. We predict that listening difficulty will increase as we decrease the SNR and detect differences in the more adverse conditions. As hypothesised in previous experiments we anticipate observing that higher quality speech synthesizers (Hybrid) will demand the least cognitive resources while low-quality speech synthesizers (Low-Quality HMM) will demand the most cognitive resources with natural speech demanding the least cognitive resources.

Noise Procedure

The same sentences that were used in the previous Experiment 2 was used for these experiments. The noise that was used is called speech-shaped noise (SSN). Speech-shaped noise is a type of noise that is designed to have a similar frequency spectrum to human speech and is typically used in research to simulate the background noise that people might encounter in real-world listening situations. The idea behind speech-shaped noise is to replicate the frequency characteristics of typical speech sounds, such as vowels, consonants, and other speech elements. This makes it a useful tool for evaluating speech quality in conditions that mimic real-world environments which helps ensure that the technology being tested works effectively in noisy environments where communication is essential. To create speech-shaped noise, the frequency spectrum and statistical properties of actual speech signals are analyzed and then synthesized to generate noise with similar characteristics. This noise can be adjusted to match different speech-related parameters, such as the spectral content, intensity, and duration, depending on the specific testing requirements. The stimuli with

speech-shaped noise used for the experiments in this thesis was created as follows:

Step 1: Analyse the spectral characteristics of typical speech signals to determine the frequency content and statistical properties.

Step 2: Use this analysis to synthesize the SSN by generating random noise that matches the spectral characteristics of speech. This can be done using various signal processing techniques, such as filtering and spectral shaping. In our work, the SSN masker was created by passing a random uniform noise through a filter with the long-term spectrum of the sentence stimuli used in the previous experiment.

Step 3: Specific SNR levels you want to create need to be selected. In our work we selected SNRs of -1 dB, -3 dB and -5 dB.

Step 4: To create varying SNRs, the level of the SSN needs to be adjusted while keeping the speech signal constant. A reference SSN level that corresponds to a specific SNR needs to be used. Then the attenuation needs to be calculated which is needed to achieve the desired SNR level by using the following formula:

Gain (in dB) = Target SNR (in dB) - Reference SNR (in dB)

Step 5: Once the attentuation is calculated, the level of the SSN needs to be adjusted by applying the calculated attenuation. This can be done using a digital audio processing tools that can be coded Matlab which is the tool used in our experiment. Repeat these steps for each desired SNR level using the reference SSN as a starting point. Once you have the various SNR levels, overlay the adjusted SSN with your speech signal of interest. This will create audio files or signals with varying SNRs.

Pre-processing

Details pertaining to the pre-processing carried out for Exp. 3 are summarized in Table 4.15. For ease of comparison we also include details for Experiment 2.

Table 4.15: Experiment analysis details of Exp. 2 and Exp. 3, including the total number of participants that were included in the analysis with the total number of recruited participants in brackets, the total number of trials remaining after the trial exclusion criteria was applied with its respective percentage shown in brackets and the mean recall accuracy percentage

Experiment	**Participants**	**No. of trials (%)**	**Mean Recall Accuracy %**
2	15(15)	869 (77)	96
3A	15(15)	886 (79)	91
3B	15(15)	849 (75)	80

Results: Intelligibility

Recall accuracy in both noise experiments were greater and equal to 80%. Therefore, even under noisy conditions intelligibility remained close to ceiling. Table 4.16 shows the WERs for each speech condition in all experiments compared. In Exp. 2, Hybrid has the lowest WER and natural has the highest. In Exp. 3A, Natural speech has the lowest WER and is significantly[11] different to all other conditions except Hybrid. In Exp. 3A and 3B, Unit Selection has the highest WER and is significantly different to all other conditions. All other speech conditions in Exp. 3B were equivalent. As expected, all WERs were higher when listening in noise at -3dB than -1dB, as listening in an increased noise environment would increase the difficulty of encoding the words being spoken especially in the instance that no speech enhancement techniques have been applied.

Table 4.16: WER percentage of speech conditions in Exp. 2 and Exp. 3

Speech Condition	**WER %**		
	Exp. 2	**Exp. 3A**	**Exp. 3B**
Natural (Human)	6	4	15
Hybrid	2	10	22
Unit Selection	4	15	29
HMM	5	8	17
Low Quality HMM	4	8	17

Results: Self-reported measures

Table 4.17: Self-reported measures (Median Score, – higher is better) and (Cognitive Load, CL – lower is better)

Speech Condition	**Exp. 2**		**Exp. 3A**		**Exp. 3B**	
	MOS	**CL**	**MOS**	**CL**	**Median**	**CL**
Natural (Human)	4	1	4	3	3	4
Hybrid	3	3	2	3	2	4
Unit Selection	2	3	2	4	2	4
HMM	2	3	2	4	2	3
Low Quality HMM	1	4	1	4	2	3

The medians of the self-reported measures for Exp. 2 and 3 are presented in Table 4.17. Across all three sub-experiments Natural speech was rated the most natural sounding. Perception of natural speech was not affected when listening in noise in -1dB SNR but dropped in -3dB SNR. As the SNR decreased, CL for natural speech increased as expected. Hybrid speech in quiet was the most natural amongst the speech synthesizers. However, when listening in noise this was no longer true. Hybrid, Unit Selection and HMM were found equally natural. In quiet and the -1dB

[11]Please note: All statistical results can be found in Appendix B.

SNR, Low-Quality HMM sounded the least natural, but in the -3dB SNR, naturalness of the Low-Quality HMM became equivalent to all other speech synthesizers. These results tell us that under more difficult SNRs, participants find it more challenging to distinguish the naturalness between the various speech synthesizers. This is not surprising - when listening under adverse conditions the primary goal is to understand what is being said and therefore our mental resources are likely to be allocated more towards the task of decoding the words rather than paying attention to how natural a voice sounds. This finding is an important one, as it suggests that when listening under difficult noise conditions subjective naturalness scores are likely to become less reliable. In other words, the naturalness scores for these speech synthesizers may not necessarily reflect the true naturalness of the system under noisy conditions as listeners struggle to perceive their naturalness especially since no speech enhancement techniques have been applied.

With regards to the self-reported cognitive load in quiet, the easiest and most difficult speech conditions (Natural and Low-Quality HMM respectively) are easy to identify. However when listening in noise, we see that distinctions between the conditions become less apparent. In the easier -1dB SNR, Natural and Hybrid are equivalent and Unit Selection, HMM and Low-Quality HMM are equivalent. In the more difficult -3dB SNR, Unit Selection becomes equivalent to Natural and Hybrid. An interesting observation is that for the two HMM conditions, CL is perceived to be easier when listening in -3dB SNR than listening in -1dB SNR.

A significant[12] negative correlation between the naturalness and cognitive load was found in Exp. 2 (corr=-0.68). In Exp. 3A, this correlation is significant but weaker (corr=-0.45) and in Exp. 3B, the correlation is not at all significant. This result suggests that when listening in quiet, listeners' are able to pay more attention to speech naturalness and therefore naturalness becomes a contributing factor when scoring CL. When listening conditions become more difficult like in the case of listening in -3dB SNR, it becomes difficult for listeners to perceive how natural the speech sounds and thus it is less likely to be considered when scoring the CL. As a result, the correlation between naturalness and cognitive load becomes weaker.

Results: Analysis of Variance

Table 4.18: ANOVA results for mean pupil dilation, peak pupil dilation and peak latency in Exp. 2 and Exp. 3

Experiment	$df1$	$df2$	**Mean**		**Peak**		**Latency**	
			F	p	F	p	F	p
Exp. 2	4	56	2.04	0.10	1.70	0.16	1.95	0.12
Exp. 3A	4	56	0.82	0.52	1.00	0.42	0.89	0.48
Exp. 3B	4	56	0.53	0.72	0.76	0.56	0.30	0.88

For the ANOVA results, as hypothesised in earlier experiments, we expect to observe a significant effect when listening in noise for mean pupil size and peak pupil dilation which should reflect at the very least that natural speech is significantly lower than synthetic speech and that the high quality

[12]Please note: All statistical results can be found in Appendix B

speech synthesizers (Hybrid) are significantly lower than the poor quality speech synthesizers (Low-Quality HMM). The ANOVA analysis results are shown in Table 4.18. No significant differences were found.

Results: Growth Curve Analysis

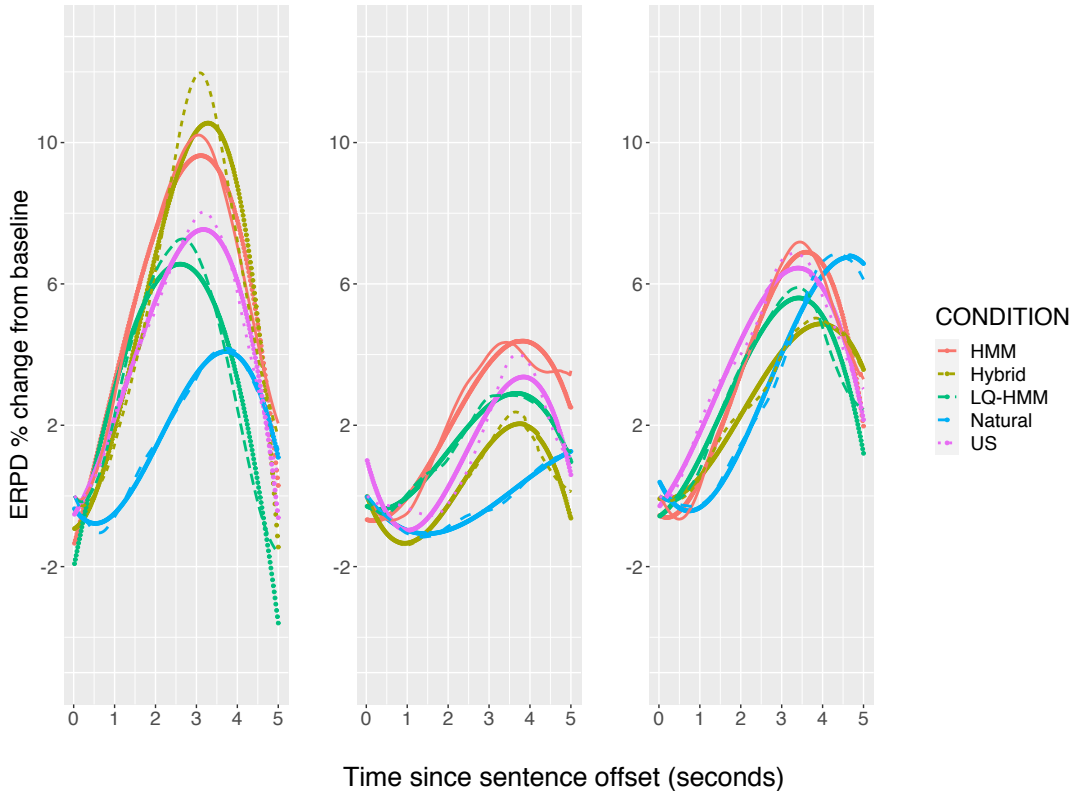


Figure 4.5: Time series line graph of the raw pupil response (dotted) and cubic model fit (solid) averaged across all participants, Exp. 2 shown on left, Exp. 3A in the middle and Exp. 3B on right.

Table 4.19: GCA parameter estimates of each time term and speech condition in Exp. 2

Condition	Intercept	Linear	Quadratic	Cubic
Natural	2.09	25.40	-6.33	-6.82
Hybrid	5.78	35.05	-42.02	-19.46
HMM	6.55	32.18	-40.49	-8.21
Low-Quality	3.99	3.77	-38.79	-0.64
Unit Selection	5.53	27.18	-31.53	-7.76

Table 4.20: GCA parameter estimates of each time term and speech condition in Exp. 3A

Condition	Intercept	Linear	Quadratic	Cubic
Natural	0.1	10.50	-5.72	-2.36
Hybrid	1.64	23.81	-6.49	-10.71
HMM	2.87	32.21	-10.29	-9.09
Low-Quality	1.47	15.61	-7.41	-6.15
Unit Selection	2.27	30.41	-3.73	-15.58

Table 4.21: GCA parameter estimates of each time term and speech condition in Exp. 3B

Condition	Intercept	Linear	Quadratic	Cubic
Natural	3.09	43.06	-4.90	-8.33
Hybrid	2.98	30.64	-5.58	-7.15
HMM	3.57	34.28	-18.20	-13.68
Low-Quality	3.43	25.21	-18.54	-8.37
Unit Selection	4.70	29.42	-22.62	-7.96

For the GCA results, we expected to see lower parameter estimates for natural speech and the high quality speech synthesizer (Hybrid) compared to the poor quality speech synthesizers. In general we expected estimates to significantly increase from Hybrid to Unit Selection to HMM to Low-Quality HMM for all time terms.

Intercept In Exp. 2 and Exp. 3A, Natural speech has the lowest mean pupil response and HMM has the highest. In both sub-experiments all conditions were significantly different to one another. In Exp. 3B, Natural and Hybrid which were equivalent has the lowest means and Unit Selection has the greatest mean and was significantly[13] different to all other conditions. An interesting observation is that all intercept parameter estimates have lower values when listening in -1dB SNR compared to listening in quiet. It is unlikely that listening effort is higher when listening under noisy conditions than in quiet and thus confirms our notion in Section 4.4.2, that the pupil response is indexing something other than listening difficulty in quiet. It is also important to note, that the initial baseline when listening in quiet is lower than the baseline when listening in noise which forces the pupil response to increase within a restricted window. Hence the change is smaller than that of listening in quiet. As expected, all intercept parameter estimates are higher in listening in -3dB than -1dB. Therefore, when listening in noise a greater mean pupil response is associated with increased listening difficulty.

Linear term In Exp. 2, Low-Quality HMM has the flattest slope and is significantly different to all other conditions. As mentioned previously, we believe this is due to low levels of engagement. In Exp. 3A, Unit Selection and HMM have the steepest slopes. Natural has the least steepest slope and was significantly different to all other conditions. This finding suggests that Unit Selection

[13]Please note: All statistical results can be found in Appendix B.

and HMM are more difficult to process in noise compared to Hybrid and Natural. This is in line with what we expect and thus confirms that we are measuring listening difficulty. In Exp. 3B, all conditions increased in steepness. Natural has the steepest slope whilst Low-Quality HMM has the least steepest slope. We observe a smaller pupil response for Low-Quality HMM in both noise conditions compared with Hybrid speech, which we know cannot be easier to process than Hybrid speech. This result therefore leads us to believe that Low-Quality HMM reaches ceiling capacity when listening in -1dB which explains the smaller pupil response.

Quadratic term In Exp. 2, Natural speech has the flattest peak and is significantly different to all other conditions. Hybrid has the sharpest peaks. In Exp. 3A, Unit Selection has the flattest peak followed by Natural speech whilst HMM has the sharpest peak. In Exp. 3B, Natural speech still has the flattest peak whilst Unit Selection has the sharpest peak. The peak shapes in Exp. 2 are all sharper than the peaks in Exp. 3A. Sharper peaks are associated with greater listening effort but this is not practical as listening in noise cannot be more difficult than listening in quiet and therefore confirms the likelihood that the pupil response is indeed not indexing listening difficulty in quiet. All speech conditions except Natural and Hybrid have sharper peaks in Exp. 3B compared to Exp. 3A. This indicates that Natural speech and Hybrid speech are processed similarly when listening in -3dB SNR and both conditions still seem manageable to listen to even in the more challenging SNR. It is interesting the Unit Selection goes from the flattest peak in -1dB to the sharpest peak in -3dB.

Cubic term In Exp. 2, Hybrid has the steepest slope and significantly different to all other speech conditions. Low-Quality HMM has the flattest. In Exp. 3A, Natural speech has the flattest slope whilst Unit Selection has the steepest slope. In Exp. 3B, all speech conditions are equivalent in slope with the exception of HMM which has the steepest slope and is significantly different to all other conditions. Unit Selection and Hybrid have smaller gradients in Exp. 3B than in Exp. 3A.

Summary

In this experiment we investigated the influence on the pupil response when listening in the presence of speech-shaped noise. Results confirmed that intelligibility is affected when listening in the presence of noise. In particular, as the SNR decreases so too does the intelligibility. With regards to naturalness perception, human speech was unaffected by the presence of noise, however in the case of synthetic speech, the perceptions of naturalness across the speech synthesizers overlap as the SNR decreases. A similar trend for the self-reported cognitive load is observed. This implies that listeners find it difficult to detect differences between speech synthesizers in a challenging noise environment. Therefore, when evaluating the naturalness of synthetic speech in the presence of noise, subjective evaluations become less discriminative. Furthermore, when listening in quiet, a strong correlation between naturalness and cognitive load is observed. However in noise, this correlation weakens as listeners' ability to distinguish differences in naturalness diminishes.

The hypothesis that the pupil response is indexing different properties when listening in quiet

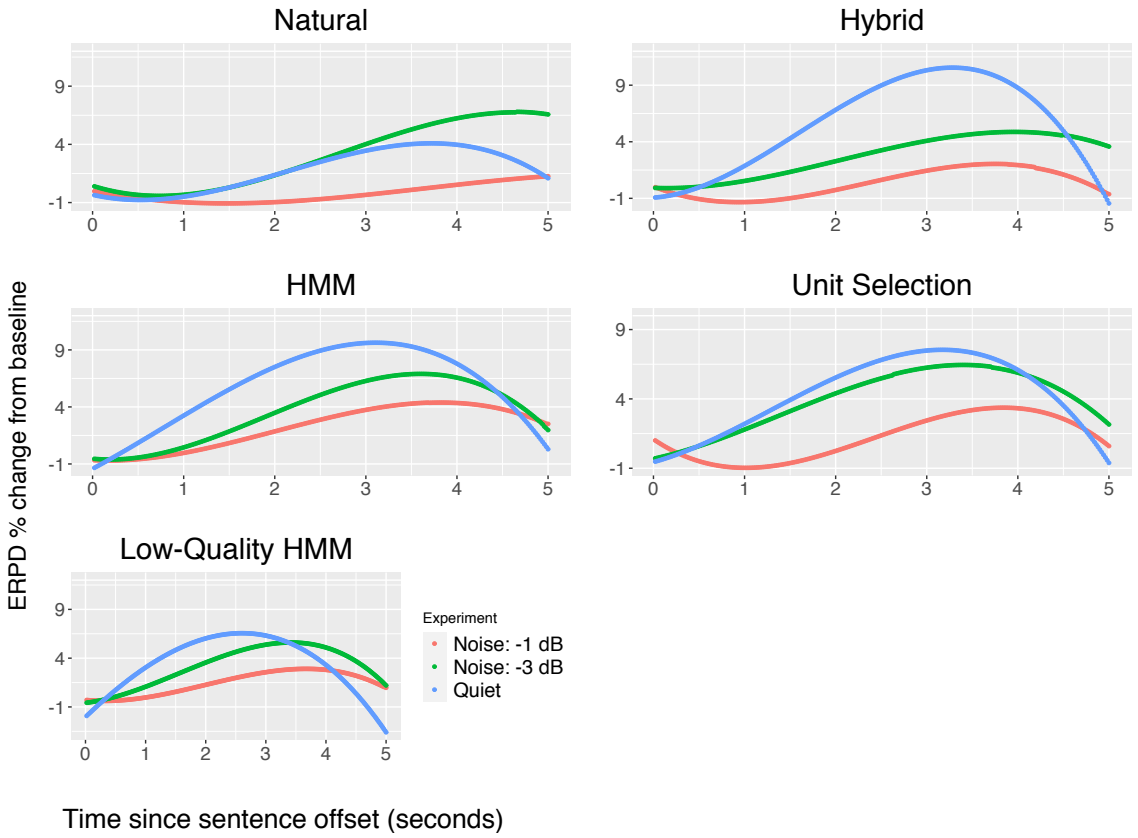


Figure 4.6: Time series line graph of cubic model fits for Exp. 2, Exp. 3A and Exp. 3B for each speech condition individually

versus listening in noise is consistently supported by the many contradictions observed in the analysis. For example, we observe in Figure 4.6 that the change in ERPD across all speech conditions when listening in quiet is greater than that when listening in noise for synthetic speech. We know for certain that listening in quiet is easier than listening in noise, yet the ERPD implies the opposite. Since the pupil response is understood to be an index of listening effort, a reasonable explanation for this contradictory observation is that in quiet, as previously explained, some other cognitive resource is being indexed other than listening difficulty such as level of engagement. In noise, the pupil response is indexing the working memory resources and therefore is associated more with greater listening difficulty. It is also important to note that lower ERPD values in noise is unexpected which could possibly be attributed to the baseline being higher in noise than the baseline in quiet which as a consequence constrains the amount of change that pupil dilation can have. Therefore, the ERPDs between listening in quiet and noise may not be comparable in this experiment.

The pupil results in the easier -1dB SNR noise condition, showed that the evoked pupil response when listening to Natural speech has the lowest and second lowest estimates across all time terms. Therefore demanding the least amount of listening effort compared to all other systems. Therefore, even in noise the listening effort of natural speech remains lower in comparison to synthetic speech. When listening to the Low-Quality HMM system, we see that all estimates are the second lowest but the peak is the second sharpest in Exp. 3A. This finding is completely unexpected as it is unlikely that Low-Quality HMM demands the second least amount of listening effort. In the literature,

such a response indicates that some resources are abandoned due to a loss in motivation and/or a listeners' willingness to perform the listening task as a consequence of being too difficult. This is an indication that for the Low-Quality HMM system it is possible that ceiling capacity is reached already in the -1dB SNR condition. Ignoring Low-Quality HMM, in Exp. 3A, Hybrid has parameter estimates that are closely follow Natural in with the exception of the slopes. Hybrid, has higher gradients for both slopes in Exp. 3A compared to natural but in Exp. 3B Hybrid and Natural have similar slopes. This suggests that natural speech is easier to process than Hybrid when listening in -1dB SNR but they appear to converge when listened in the -3dB SNR. In Exp. 3A, Unit Selection, has the highest and/or second highest estimates across all time terms except in the quadratic term where it has the flattest peak. This differs slightly to what we observed in Exp. 1, where all time terms for Unit Selection were high. It is strange that Unit Selection has the flattest peak amongst all systems. In Exp. 3B, we see high estimates for Unit Selection in all time terms but the gradient of it's rising slope is less than all other conditions except Low-Quality HMM. From all the speech synthesizers, we observe the greatest increase for Unit Selection from the -1dB to -3dB. This could explain the significant increase in peak sharpness between the two noise experiments. Therefore, Unit selection demands medium listening effort in -1dB but high listening effort in -3dB. In -1dB, HMM demands the most listening effort as all parameter estimates for HMM are the highest across all time terms. In -3dB, Unit Selection surpasses it but HMM doesn't seem far behind.

The main takeaways from this experiment is that when investigating the impact speech quality has on the pupil response, it is important to consider changes in the pupil response by examining differences from decreasing the SNR. Presumably, listening effort should increase with decreased SNRs. If the properties (mean, slopes and peak) behave differently to the expected trend of continual increase, this is then an indication that some other contributing factor is being indexed. For examples, findings in our experiments found that some other cognitive resource is indexed when listening to SMS in quiet. Findings in our experiment also showed that ceiling capacity as a consequence of fatigue is indexed when listening to degraded quality (such as speech produced by Low-Quality HMM) which resulted in a smaller pupil response. Listening difficulty is indexed in the presence of noise. Furthermore, the worst quality speech synthesizer appear to reach ceiling at -1dB whilst high quality speech synthesizers were still manageable to listeners in -3dB SNR. Therefore, it is important to consider all properties reflected by the pupil response in order to understand and interpret results and how they relate to listening effort. In conclusion, measuring the listening effort in the presence of speech-shaped noise proves viable in detecting listening effort differences between speech synthesizers and human speech as well as between each other.

Chapter 5

Summary of investigations in Part I

In Part I of this thesis, the aim was to answer the research question "How can we measure the cognitive load when listening to synthetic speech?" This research question needed to be answered with respect to determining the most suitable measurement of evaluating the cognitive load imposed on listeners when listening to speech produced by a text-to-speech synthesizer.

The two methods investigated were the dual-task paradigm as a behavioural method discussed in Chapter 3 and pupillometry as a physiological method discussed in Chapter 4. Both selected methods were chosen on the basis that they are common, non-intrusive, inexpensive and easily-accessible methods. In addition, self-reported measures were collected in support of each of the selected behavioural and physiological methods. The key findings for both methodologies are discussed in the sections that follow.

5.1 Discussion

5.1.1 Dual-task paradigm

The dual-task paradigm consisted of a sentence recognition task as the primary task and a visual-motor task as the secondary task. Difficulty of the primary task was varied by the quality of the speech listened to, where poor quality speech was expected to be harder to listen to than high quality speech. Differences in performance in the secondary task when listeners prioritise the primary task reflect a shift in cognitive resources allocated to the primary task. This interpretation assumes that: (1) performance on the primary and secondary tasks requires the allocation of cognitive resources to each task, and (2) cognitive resources are limited. Two visual-motor secondary tasks were compared, a Digit Task in Exp. 1 and a Word Task in Exp. 2. The only significant difference found in the results was between Natural and Hybrid speech when listeners performed the Digit Task. The RTs when listening to Natural speech were significantly slower than the RTs when listening to Hybrid speech. This result was surprising, as studies that have applied the dual-task paradigm in the past have shown that RTs are slower when listening to synthetic speech compared to natural speech. Therefore, previous work implied that synthetic speech is more difficult to process than natural speech [Sonntag et al., 1998]. However, the results in our work show the

opposite. If this paradigm is measuring listening effort, this would suggest that synthetic speech is easier to listen to than human speech, which seems unlikely. Therefore this leads us to believe that some other measurement is being indexed by the dual-task paradigm in our experiments. The notion of some other measurement being indexed is supported by the self-reported measures which show that Natural speech is easier to listen to than Hybrid speech. Upon further investigation, we observed that 50% of participants responded faster in the dual-task than they did in the secondary task in isolation. Therefore, deterioration in the secondary task during the dual condition was not always achieved. This leads us to believe that both the primary and secondary digit task were not cognitively demanding enough. In other words, both the primary task and secondary digit task were manageable to perform concurrently without exerting a forceful deterioration on the secondary task. Therefore, it is likely that both the primary and secondary digit task were performed comfortably using the total resources available resulting in no compromised performance on the secondary task. As a consequence, it is unclear whether these results are in fact a real indication of listening effort. In addition, it was difficult to know for certain whether the listener prioritised the primary task over the secondary task despite being instructed to do so. This uncertainty poses a challenge in the interpretation of the results. If 50% of the RTs were faster when participants performed the secondary task in dual, compared to performing the same task in isolation, then it is possible that some listeners did not prioritise the primary task. Another possibility is that listeners naturally performed better in the dual-task as a consequence of the training effect. These uncertainties made it difficult to interpret the results.

We attempted to increase the load by using words instead of digits. Using words instead of digits meant that listeners would likely be utilizing the same cognitive resources, as both the word and sentence recognition tasks both require linguistic knowledge. However, no significant differences in the word task were found. The RTs when performing the word task were only marginally slower than the RTs when performing the digit task for all speech synthesizers. This implies that the load demanded by the digit task and word task did not significantly change as we had hoped for. Determining the impact of these results is difficult because listening effort quantified by this method is inferred indirectly from change in secondary task performance. This, however, is only applicable if there is a deterioration in performance in the secondary task. Furthermore, the self-reported cognitive load measures collected during these experiments did not correlate with the RTs, but were meaningful in that they showed that cognitive load appears to be negatively correlated with naturalness. Perceptually, listeners were able to differentiate between the cognitive effort expended when listening to the various speech synthesizers on the basis of how natural-sounding the speech was. When the word task was used as a secondary task, this differentiation was clearer than when the digit task was used. In the digit task the high quality speech synthesizers were perceived to be equally easy to listen to whilst in the word task, Hybrid speech (the highest quality speech synthesizer) was found to be the easiest to listen to. In other words, when performing the word task as the secondary task, the highest quality TTS synthesizer was perceptually teased apart from the rest. Therefore, to detect differences in cognitive load between speech synthesizers it is important that the loads imposed by both primary and secondary tasks are great enough. Unfortunately, this

was not successfully achieved in our dual-task experiments which were found to be unreliable and did not provide meaningful results for measuring the cognitive load of synthetic speech.

5.1.2 Pupillometry

In our experiments, we were particularly interested in listening effort, so listeners' pupil sizes were collected whilst they listened to audio samples produced by human speech and four speech synthesizers.

Since the pupil response is involuntary, it is sensitive to many factors that need to be carefully controlled during experiments in order to obtain meaningful and accurate results that we desire. For the purpose of indexing listening effort, factors such as those discussed in Section 4.2 need to be taken into consideration. In addition to those factors, the type of sentence material can influence changes in the pupil response. In traditional tests used to evaluate TTS, SUS are used when evaluating intelligibility and SMS are typically used when evaluating naturalness. However, when listening to SUS sentences compared to SMS, the cognitive processes allocated to understanding the speech may differ. The intent is to be able to index the amount of effort exerted which is representative of a scenario when listening to TTS in the real world. Listeners do not listen to SUS in the real world and therefore such a technique is ecologically invalid. Despite SUS being ecologically invalid, it is important to compare the results of SUS and SMS out of concern that the overall load of listening to SMS would be insufficient to evoke significant differences. Listening to SMS in quiet may be effortless. If, when listening to SMS, a pupil response isn't great enough to detect differences, then the load of listening needs to be increased. This can be achieved by the listener performing the listening task in the presence of noise. For these reasons, three main experiments were conducted in Chapter 4. The first investigated how the pupil response is influenced by listening to SUS, the second compared how the pupil response changes when listening to SMS compared to SUS and the third investigated the influence on the pupil response when listening in the presence of speech-shaped noise.

To our knowledge, ours is the first attempt to measure the listening effort of synthetic speech using pupillometry. Therefore, it was important to compare the results obtained in our experiments against the results obtained in the Blizzard Challenge as a means to validate our findings. In terms of naturalness, the medians reported in the 2010 Blizzard Challenge are similar to the medians obtained in our experiment (Table 4.5 Exp.1B). Unit Selection, however, was rated more natural than the HMM models in the Blizzard Challenge, whilst in our experiments, Unit Selection and the HMM models were equivalent. In the 2011 Blizzard Challenge, Natural was rated the most natural and all other systems were equivalent. In our experiment (Table 4.5 Exp.1A), Natural was also rated the most natural but the Low-Quality HMM was rated the least natural and Hybrid was rated the most natural from all the speech synthesizers. Although, both experiments have small differences, the overall ranking between best to worst remains the same. In terms of intelligibility, the WERs in our experiments were much lower than the WERs in both Blizzard Challenges. Natural speech had the lowest WERs in all experiments and Unit Selection the highest in Blizzard 2010 and Low-Quality HMM had the highest in Blizzard 2011. Once again, despite the WERs being lower,

the ranking between systems remained the same. These rankings were thus used as benchmarks. If cognitive load is influenced mostly by how natural the speech sounds then intuitively as naturalness of synthetic speech reduces the cognitive load should increase accordingly. Similarly, if cognitive load is influenced by intelligibility then as intelligibility of synthetic speech reduces the cognitive load should increase accordingly.

Self-reported cognitive load

For the synthesizers selected from the 2011 Blizzard Challenge, Hybrid was perceived to be the easiest TTS system to listen to and the Low-Quality HMM was the most difficult to listen to. For the synthesizers selected from the 2010 Blizzard Challenge, Hybrid was perceived to be the easiest TTS system to listen to and Unit Selection was the most difficult to listen to. The self-reported cognitive load measures aligned with both the naturalness and intelligibility scores. This implies that self-reported cognitive load corresponds well with results obtained from traditional naturalness and intelligibility evaluation methods (as discussed in Section 2.3). If intelligibility is a contributing factor to cognitive load then this would mean that traditional intelligibility tests are sufficient to tell us how effortful synthetic speech is to listen to. However, in our analysis, intelligibility was controlled by taking into consideration the trials that were intelligible. Yet, differences in the pupil response were still observed. Therefore, implying that the self-reported cognitive load measurement is more than just the intelligibility. We observed that a significant negative correlation between the naturalness scores and the self-reported cognitive load scores were found. This implies that naturalness scores do have an influence on the self-reported cognitive scores. Although self-reported measures give us an indication of which TTS systems are more difficult to listen to, they are not sufficient to help us understand *why* it is more effortful to listen to. This motivates extending to new measurements like pupillometry that have the potential to provide much more meaningful insights than a 1-dimensional score obtained in a self-reported assessment.

Growth Curve Analysis

As discussed in Section 4.3.2, Growth Curve Analysis was applied. A cubic model was fitted to the curve produced by the raw pupil size data collected over a time interval when a listener listened to synthetic speech. From this data, we can estimate how effortful it is to listen to the speech by analysing the properties of the curve as shown in Table 4.1. Generally, high listening effort is associated with a large mean pupil dilation, steep rising slope, sharp peak shape and steep falling slope. In contrast, low listening effort is associated with low mean pupil dilation, broad peak shape and rising and falling slopes that have low gradients.

Experiment 1: SUS In Exp. 1A, Natural speech evoked a pupil response that has the lowest mean and broadest peak in both experiments. Similarly, Hybrid had the second lowest mean and peak. Natural and Hybrid speech was therefore shown to demand the least amount of listening effort. Interestingly, Natural and Hybrid speech did not have the lowest slope gradients which

contradicts the properties we typically associate with low listening effort. Unit Selection has the greatest mean, steepest slopes and the second sharpest peak. All of which we can associate with high listening effort. This finding was interesting, as Unit Selection was not rated the lowest in naturalness nor did it have the highest WER. Low-Quality HMM, the worst rated synthesizer and highest WER, has the second greatest mean and sharpest peak both of which are associated with a high listening effort. However, in the rising slope, it has the second lowest gradient. Similarly, HMM has a higher mean and sharper peak than Natural and Hybrid, yet it's slopes gradients are both the lowest of them all. Therefore, Natural and Hybrid speech have steeper slopes than both the HMM models. It seems likely that the linear and cubic terms in this experiment is indexing something else other than listening difficulty.

In Chapter 2, we mentioned that research investigating listening effort identified four dominant cognitive processes for speech processing ie., working memory, attention, processing speed and linguistic knowledge. In our experiments, linguistic knowledge was controlled by balancing all sentences, systems and participants. Therefore linguistic knowledge could not have influenced differences in the pupil response within an experiment. Working memory is the process that is related to listening difficulty, if speech is more difficult to process, the working memory works harder to process the speech. Both Natural and Hybrid speech have steeper slopes than the both HMM systems which we know is unlikely to be easier to process than human speech and the highest quality speech synthesizer. Therefore, we are led to believe that working memory is not the resource being indexed in these terms. If attention resources are being indexed instead, this could relate to level of engagement. Attention resources indexed as a result of high levels of engagement can be seen as both positive and negative. Positive in that listeners choose to allocate more attention to listen to speech as a form of engaged interest and motivation, and negative in that listeners are forced to pay closer attention to speech that is challenging to process. From the positive point of view, this implies that the more natural sounding speech (natural and Hybrid) evoked a steep slope in comparison to the HMM systems as it was more engaging to listen to. From the negative point of view, this could explain why Unit Selection, which was found to demand more listening effort in other terms, also evoked a steep slope. Another plausible resource that is being indexed could be speed of processing. If Hybrid and Natural speech are easier to process than the HMM systems, then it is likely that they are processed faster and therefore the gradients are steeper. However, if the slope was indexing speed of processing alone then Unit Selection, which is harder to process shouldn't have elicited the steepest slope. Therefore, based on these theories, we are leaning towards attention resources being indexed by these terms. It is out of the scope of this thesis to understand which exact resource is being indexed. At this stage we are only interested in listening difficulty and it is clear enough that listening difficulty is not being indexed by these terms in this experiment.

In Exp. 1B, we observe similar results. Natural and Hybrid speech again have the lowest means and broadest peaks but both have steeper slopes than both HMM systems. This confirms that the slopes appear to be indexing something other than listening difficulty. Unit selection in this experiment also evokes the greatest mean, second highest steepest rising slope, sharpest peak and steepest falling slope. All of which are associated with high listening effort. This result aligns with

the naturalness and intelligibility scores as Unit Selection was rated one of the poorest systems and has the highest WER. The fact that Unit Selection performed the worst across both experiments, despite being different versions of unit selection models, using different sentence material and a different speaker suggests that the general architecture of Unit Selection is the most difficult speech synthesizer to listen to. Given that this result did not align with the traditional evaluation methods in Exp.1A, proves that traditional methods alone are not sufficient enough to measure the cognitive load of synthetic speech. Physiological measures have the potential to provide us with a deeper understanding of how TTS interacts with the human cognitive processing system beyond the two traditional evaluation methods. Therefore, this first experiment provided sufficient evidence that pupillometry, as a physiological measure, could be a potential method for measuring the listening effort of synthetic speech. This method proved useful in detecting differences between various TTS systems. However, we were concerned about the ecological validity of the results as SUS are not representative of speech processing in the real-world.

Experiment 2: SMS The second experiment aimed to address the concern of the ecological validity when using SUS by replacing SUS with SMS. When listening to SUS, Natural was perceived easy to listen to. Low-Quality HMM was the most difficult to listen to. When listening to SMS, listeners found it more difficult to differentiate between the different systems. Natural was perceived the easiest and Low-Quality HMM was perceived as the most difficult but Hybrid, Unit Selection and HMM were all equally difficult to listen to. This indicates that when we process meaningful sentences, perceptually our cognitive processing system compensates for differences in speech quality that lie between completely unnatural (for eg., Low-Quality HMM) or completely natural (for eg., human speech) which makes it harder to detect differences between such systems.

When listening to SMS produced by natural speech, the pupil response has the lowest or second lowest parameter estimates across all time terms. Therefore, natural speech demanded the least listening effort which is intuitive as natural SMS are effortless to process. When listening to SMS produced by the Hybrid synthesizer, the pupil response has either the highest or second highest parameter estimates across all time terms. This implies that Hybrid demands the most listening effort. In contrast, when listening to Low-Quality HMM, the pupil response has lower parameter estimates than Hybrid in all terms except the peak. It is unlikely that the worst quality TTS system demands less resources than the best quality TTS system. Therefore, this provides evidence that the pupil response appears to be indexing something other than listening difficulty, in the case of synthetic speech only. Therefore, the type of sentence material is important when evaluating the cognitive load of synthetic speech and indicates that natural speech is processed differently to synthetic speech. Processing of natural speech is unaffected by sentence material which is expected as natural speech is considered to be effortless. Synthetic speech, however, does demand more mental resources when listening to SMS than natural speech but at this stage we do not know for certain whether these additional resources will lead to negative implications. Therefore, when measuring the cognitive load of synthetic speech, we first need to ask ourselves what the main purpose is for measuring cognitive load. If the purpose is to optimise cognitive load of synthetic speech in relation

to developing TTS systems that demand the same resources as natural speech, then SMS in quiet conditions does not seem to be a suitable technique as they are indexing different resources and by having different metrics this makes it difficult to know for certain when these metrics - if ever - will converge.

Experiment 3: Noise The third experiment was an alternative approach to measure the cognitive load of synthetic speech by introducing background noise. The idea was that by adding speech-shaped noise, sufficient load will be placed on the listener such that the pupil response indexes more the allocation of resources to the working memory and not any other cognitive processes. Thereby, indexing only listening difficulty. Listening in the presence of noise is also considered to be more ecologically valid than listening in quiet conditions, especially as TTS becomes embedded in more real-world applications and therefore listening is more likely to take place in changing environmental conditions. Speech-shaped noise is not entirely representative of the real-world conditions but it is a starting point. For now, the purpose was merely to investigate how the pupil response behaves when listening to synthetic speech in the presence of noise. The self-reported measures for cognitive load showed that listeners found it difficult to detect differences between speech synthesizers in the presence of noise. When synthetic speech was evaluated in the presence of noise, the self-reported measures became less reliable. This motivates the need to explore new evaluation measures that use physiological measures like pupillometry that have the potential to provide more meaningful results. Larger pupil dilations were evoked when listening to speech in the -3dB SNR compared to the -1dB SNR but even larger pupil dilations were evoked when listening in quiet. Since we know listening in quiet can not be harder than listening in noise, this result confirms our notion that in quiet the pupil response is more likely to be indexing something else, and in noise it is more likely to be indexing working memory resources which is the measurement we desire.

In the -1dB SNR, HMM and Unit Selection both have parameter estimates with the highest listening difficulty whilst Natural has parameter estimates with the lowest listening difficulty. Hybrid has a mixture of parameter estimates that demand medium to high listening difficulty but leans more towards medium listening difficulty. Low-Quality HMM has conflicting properties. It is unusual that the Low-Quality HMM is not entirely associated with high listening difficulty when it was specifically chosen in this experiment as being the speech synthesizer with the poorest speech quality. Pupillometry studies have reported that when a signal over-exerts the mental processing system it is likely to elicit a smaller pupil response [Wagner et al., 2016]. We believe this is what has happened in the case of listening to Low-Quality HMM and thus explains the conflicting properties observed in the results. With the exception of Low-Quality HMM, all other speech conditions follow the expected trend in terms of listening difficulty with HMM and Unit Selection as the most difficult synthesizer to listen to and Hybrid the easiest synthesizer to listen to.

In the -3dB SNR, all speech conditions have higher means than the highest mean observed in the -1dB SNR - Unit Selection having the highest mean of all. If both HMM conditions had reached ceiling at the -1dB SNR, then it intuitively makes sense why the Unit Selection condition elicits the highest mean. However, considering all parameter estimates of HMM, we see that these correspond

to high listening difficulty with little evidence indicating signs of fatigue like we observed for the Low-Quality HMM. As observed in the -1dB noise condition, Unit Selection also has parameter estimates for all other terms that indicate high listening difficulty. Therefore, Unit Selection and HMM both demand high listening difficulty in the -3dB SNR condition. Hybrid and Natural behave similarly. We also observed that the difference in ERPD between the -1dB and -3dB for Natural is larger than Hybrid. The behavior of natural speech in the -1dB seems strange and therefore it is questionable whether sufficient load has been placed on the listener in -1dB noise condition. In the -3dB noise condition, Natural and Hybrid speech are equivalent and demand the same amount of listening effort.

Overall, we observe that Hybrid demands more cognitive load than Natural but doesn't seem too far behind. HMM and Unit Selection interchangeably demanded the most cognitive resources in both noise conditions whilst Low-Quality HMM was difficult in both noise conditions and reached ceiling capcity in the easier noise condition. These findings provide meaningful evidence of differences in processing various types of speech in the presence of speech-shaped noise and therefore proved to be a viable method for measuring the cognitive load of synthetic speech. However, when investigating the impact speech quality has on the pupil response, it is important to consider changes in the pupil response by examining differences by decreasing the SNR, as this tells us the upper bound (maximum SNR that listeners can manage when listening to) each synthesizer in terms of a listeners' capacity to successfully process it. These experiments did not reveal the upper bound for natural speech and therefore in later experiments, the SNR noise level investigated in the experiments in the chapters that follow will include a -5dB SNR condition.

5.2 Concluding remarks

Applying the dual-task paradigm in our work, we did not obtain significant results. Applying pupillometry, however, we obtained significant and meaningful results. The pupil response, when listening to both human speech and synthetic speech, was sensitive to changes in speech quality and therefore it was possible to detect significant differences between the various speech synthesizers compared. In addition, pupillometry is an online measure that is able to provide a real-time measurement whilst listening is actively taking place. The dual-task paradigm, however, is not online and a response is collected at the end of the listening task. Furthermore, a drawback of the dual-task is that it is based on the limited capacity assumption and therefore it can only provide meaningful results when both tasks are performed in dual demand resources that exceed cognitive capacity. Perhaps, making the primary task one where the listener listens in the presence of noise may exert a sufficient load to achieve more reliable results in future work. The pupil response is an involuntary response and can therefore be considered as more reliable, provided the listener is motivated and willing to perform the task and a sufficient amount of load is placed on the listener. If insufficient load is placed, we instead observe that the pupil response indexes some other cognitive process other than listening difficulty which too could be meaningful for other investigations. However, since we are interested in listening difficulty in this thesis, the most viable approach is to setup

the experiments in the presence of noise. Therefore moving forward, pupillometry was the chosen method for measuring cognitive load of synthetic speech. In Part II of this thesis, all experiments will adopt the procedure followed in the second and third experiments that investigate the influence of the pupil response when listening to SMS in quiet and in the presence of noise.

Part II

Using pupillometry to measure the cognitive load of state-of-the-art TTS

Recap

In Part I of this thesis, we investigated two methods for measuring the cognitive load of synthetic speech, the dual-task paradigm discussed in Chapter 3 and pupillometry discussed in Chapter 4. In Chapter 5, we concluded that pupillometry was the more reliable method and proved to be sensitive to detecting differences in cognitive load between human speech and synthetic speech as well as between the various speech synthesizers compared. In Part II, pupillometry will therefore be the method applied in all experiments.

In addition the second research question "Does synthetic speech demand greater cognitive effort compared to human speech" was answered. The key findings reported in Part I with respect to the cognitive load of synthetic speech was that synthetic speech is harder to listen to than human speech. In addition, results showed that, Hybrid speech synthesizer - the best quality TTS system compared in our evaluations - demanded the least listening effort amongst all speech synthesizers compared whilst Unit Selection demanded the most. The Low-Quality HMM synthesizer was the poorest speech synthesizer and results showed evidence that suggest that listeners reached ceiling capacity by evoking a smaller than expected pupil response. According to previous work [Winn et al., 2018], such a response occurs when a task is too difficult resulting in listeners experiencing a lack of motivation and/or willingness to listen. Whilst these results were meaningful in determining the validity of the method and indicated which synthesizers are generally better than others, the results did not provide us with much in-depth knowledge as to *why* cognitive load was high.

Therefore, in addition to applying pupillometry to state-of-the-art TTS systems in Part II, we also aim to understand the contributions to an increased cognitive load by answering the research question "What are the contributing factors that lead to increased cognitive load of synthetic speech?" In Chapter 6, we investigate contributions of vocoder speech parameters in a conventional DNN-based SPSS system as described in Chapter 2, Section 2.1.2. In Chapter 7, we measure the cognitive load of state-of-the-art models such as sequence-to-sequence based speech synthesis as described in Chapter 2, Section 2.1.3. Finally, in Chapter 8 we discuss and summarize the key findings of all experiments conducted in Part II.

Chapter 6

Contributions of DNN-based speech synthesis

In this chapter, we measure the cognitive load of a DNN-based speech synthesis system but we take it a step further than we did in the experiments conducted in Part I. We do this by investigating the contributions to cognitive load of each vocoder speech parameter modelled in the chosen DNN system. The aim of this experiment is to understand which speech parameters modelled within a DNN-based speech synthesis system contributes most to an increased cognitive load. Understanding these contributions are important as it gives us an indication as to which properties of the signal need to be optimised, which can suggest new ways to improve TTS that is capable of generating speech that has reduced cognitive load.

We start this chapter by describing the specific DNN speech synthesizer architecture that was evaluated in this work. We then describe our methodology undertaken to produce the various forms of synthetic speech that were necessary to understand the contributions of each vocoder speech parameter. This is followed by an overview of the experimental setup and presentation of our results. Finally, we conclude this chapter by summarizing the key findings.

(This chapter expands Govender et al. [2019a])

6.1 Introduction

In Part I of this thesis, we concluded that pupillometry is a reliable method for measuring the cognitive load of synthetic speech where the properties of the pupil response is capable of indexing the extent of listening effort exerted when listening to various forms of speech. Based on our observations, it was hypothesised that depending on the listening environment, such as listening in quiet, the pupil response appears to be indexing levels of engagement and/or attention. When listening in noise, the pupil response appears to be indexing listening difficulty. Results showed that even the best quality TTS synthesizer (evaluated in Chapter 4) demanded high listening effort as signal-to-noise ratios (SNRs) decreased whilst human speech, in easier SNRs conditions demanded low listening effort but converged with the highest quality speech synthesizer in more challenging

SNRs.

Recently with increased processing power, neural networks have become popular and in most cases DNN models have replaced HMM models in SPSS which has lead to a rapid improvement in the speech quality of TTS synthesizers. Therefore, the results obtained in Part I, are not a reflection of listening difficulty of the current-state-of-the-art models. Therefore, the aim of the work presented in this chapter is to investigate whether synthetic speech produced by a DNN-based speech synthesizer still demands greater listening effort than human speech. The work presented in this chapter to our knowledge, is the first to evaluate the cognitive load of a DNN-based speech synthesizer. In addition to this, we delve deeper by producing various forms of synthetic speech produced by the same DNN-based speech synthesizer. This is achieved by producing synthetic speech that simulates stepping gradually from human speech to synthetic speech (more detail to follow in the next section). The motivation of producing these gradual steps between human and synthetic speech is to help us in gaining a better understanding as to which properties of synthetic speech demand an increased cognitive load compared to that when listening to human speech.

6.2 Methodology and Implementation

Data 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 2542 sentences were used. For training, validation and testing 2072, 200 and 270 sentences were used respectively. Since we wish to measure cognitive load of synthetic speech for real applications, semantically meaningful sentences (SMS) are used in all experiments presented in this chapter. The sentences were taken from the Glasgow Herald newspaper. Test sentences can be found in Appendix C.

DNN-based synthesis system In Chapter 2, we introduced DNN-based speech synthesis and provided a high-level overview of the architecture of the Merlin SPSS system, which is the system built in this chapter. To summarize, a duration and acoustic model is trained separately. The duration model is trained using binary linguistic features produced by a front-end. Merlin requires the use of an external front-end. The front-end used in our experiments is Festival [Black et al., 1998] which produces a linguistic specification that is converted to a format in Merlin which complies with HTS-style labels with state-level alignments [Wu et al., 2016]. Merlin converts this linguistic specification into binary and continuous features that are used as inputs for the training of both the duration and acoustic models. Prior to training the duration model, forced alignment of the raw speech and text needs to be performed to obtain natural durations which are used as targets when training the duration model. Merlin performs the forced-alignment using a HMM-based forced aligner. Prior to training the acoustic model, vocoder speech parameters are extracted from the raw speech data. In our experiments, the WORLD vocoder was used [Morise et al., 2016]. The speech parameters extracted for WORLD include 60-dimensional mel-cepstral coefficients (MCC), 25 band aperiodicities (BAPs) and logarithmic fundamental frequency (log F0) at 5 ms frame intervals. These are used as targets for the acoustic model training. At synthesis time, text is first converted into a

binary linguistic representation using the front-end and this representation is passed as input to the already trained duration model which predicts the duration. Similarly, the same binary linguistic representation is passed as input to the already trained acoustic model which predicts the vocoder speech parameters. A parameter generation algorithm within Merlin, subsequently generates the vocoder parameters corresponding to the predicted durations, which are then used by the vocoder to reconstruct a speech waveform.

Model training The standard DNN configuration recipe in Merlin called "build your own voice" [1] was used to build the various DNN systems. The acoustic and duration models comprised of 6 feed-forward hidden layers; each hidden layer has 1024 hyperbolic tangent units.

Audio samples To better understand the contribution of each parameter, we constructed a range of systems that varies from copy-synthesis (i.e., vocoding) to a fully trained DNN TTS model as was just described. In each of the remaining built systems, some speech parameters (e.g., spectral envelope) are combined with speech parameters from copy-synthesis (e.g., F0) which is predicted from text. Table 6.1 shows the various systems built in terms of their configurations. System B is constructed using copy-synthesis which reconstructs the speech from the speech parameters extracted from the raw speech data. Therefore, the influence on the pupil response when listening to System B, simulates the contribution to increased cognitive load as a result of any loss in quality introduced by the vocoder alone. System F is a fully trained DNN model without any modifications and therefore evaluates the contribution to increased cognitive load demanded when listening to a DNN TTS system. Systems C and D combine spectral parameters from copy-synthesis with F0 parameters predicted from text and vice versa. These systems simulate the contributions of predicted F0 in System C and predicted MCC in System D in the context that that F0 and MCC are independent. In other words, F0 and MCC features are predicted separately. System E combines predicted spectral and F0 features using durations copied from human speech recordings by forced alignment. System E, therefore simulates the contributions of predicted duration to increased cognitive load. All stimuli were additionally mixed with speech-shaped noise at SNRs -1dB, -3dB and -5dB, chosen such that the cognitive load is increased whilst intelligibility remains close to ceiling.

Set-up The same pupillometry set-up as described in Section 4.2, is used for the pupil data collection in this chapter. In summary, the speech stimuli were played to listeners through headphones in a noise-and light-controlled room. Simultaneously, the pupil size was measured using an eye tracker.

Presentation Audio stimuli were presented in 6 blocks plus a practice block at the start of the experiment. Blocks were arranged using a 6 x 6 Latin square design to ensure all listeners, systems and sentences were equally balanced. The practice block comprised of 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 blocks. Each of the subsequent blocks had 20 trials. All sen-

[1] `https://github.com/CSTR-Edinburgh/merlin/tree/master/egs/build_your_own_voice`

Table 6.1: 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

tences within each block were randomized except the first 5 sentences which were kept fixed across participants as they were discarded during the analysis. At the end of each block, self-reported cognitive load and naturalness scores were collected on 5-point rating scales (1 - Very Unnatural, Very Difficult; 5 - Very Natural, Very Easy).

Participants Participants were recruited from university students and staff, ranging in age from 19 to 37 years. All participants were native English speakers with no self-reported hearing problems.

Pupil data processing, **Trial Exclusions**, **Baseline Correction** and **Post-processing** in this chapter all followed the same procedures as described in Section 4.2.

Analysis The peak picking analysis did not prove to be meaningful in Chapter 4 and therefore only Growth Curve Analysis (GCA) as described in Section 4.3.2 was used to analyse and interpret the pupil data in this chapter.

6.3 Experiments

This chapter comprises of four experiments that aim to (1) investigate the cognitive load of speech produced by a DNN-speech synthesizer in comparison to human speech and (2) understand which vocoder speech parameters contributes to an increased cognitive load. Each experiment investigates these aims under one specific listening condition. Experiment 1 investigates the influence on pupil response when listening to human speech and synthetic speech produced by the various configurations (described in Table 6.1) in quiet. Experiment 2 investigates the influence on pupil response when listening to human speech and synthetic speech produced by the same configurations in noise. Although we found that listening in quiet appears to be indexing something else and not listening difficulty, we included it here as a way to validate whether those findings hold true in a separate experiment. Our hypotheses for this experiment is that a full DNN TTS speech synthesizer will demand more cognitive resources than human speech when listening in noise but the gap would be smaller than what was observed in previous experiments for a Hybrid speech synthesizer. In quiet, we expect to see the reverse, whereby natural speech would demand more resources than the

full DNN TTS speech synthesizer which will validate our findings of the previous experiments. In terms the configurations, in quiet, we expect to see parameter estimates decrease as we step from System A to F, whilst in the noise experiments, we expect to see parameter estimated increase. In addition, we hypothesise that the contributions to an increased listening difficulty stem from a poor MCC and duration prediction in a full DNN TTS system as well as the vocoder being evaluated in this experiment which is an old tradition vocoder model.

6.3.1 Experiment 1: Quiet condition

Pre-processing

Details pertaining to the pre-processing carried out for Exp. 1 are summarized in Table 6.2.

Table 6.2: Experiment analysis details of Exp. 1, including the total number of participants that were included in the analysis with the total number of recruited participants in brackets, the total number of trials remaining after the trial exclusion criteria was applied with its respective percentage shown in brackets and the mean recall accuracy percentage

Experiment	**Participants**	**No. of trials (%)**	**Mean Recall Accuracy %**
1	24(24)	1909 (88)	96.5

Results: Intelligibility

Recall accuracy in Exp. 1 is greater than 90%. Therefore, intelligibility for speech produced by a DNN-based speech synthesis is close to ceiling. Table 6.3 shows the WERs for each system (refer to Table 6.1) compared. Human speech and System D (Predicted MCC, Vocoded F0) has the lowest WER whilst the full DNN TTS system has the highest WER. However, no significant[2] differences were found between any of the systems evaluated in terms of their WERs. Therefore, if any cognitive load differences are observed, this will indicate that the cognitive load differences are a result of some other contributing factor other than intelligibility.

Table 6.3: WER percentage of each speech system in Exp. 1

System	**WER %**
A (Human)	2
B (Vocoded)	4
C (Predicted F0)	4
D (Predicted MCC)	2
E (Natural Duration)	3
F (full DNN)	5

[2]Please note: All statistical results can be found in Appendix D.

Results: Self-reported measures

Table 6.4: Self-reported measures (Naturalness Score, – higher is better) and (Cognitive Load, CL – lower is better)

System	**Exp. 1**	
	Naturalness	**CL**
A (Human)	4	1
B (Vocoded)	4	2
C (Predicted F0)	3	2
D (Predicted MCC)	3	2
E (Natural Duration)	2.5	2.5
F (full DNN)	3	3

The medians of the self-reported measures for Exp. 1 are presented in Table 6.4. Human and vocoded speech was rated the most natural sounding. Human speech was found to be significantly[3] different to all TTS system configurations excluding System B. Therefore, listeners could distinguish between human speech and synthetic speech. Vocoded speech was only configuration found significantly different to System E and F. It is not surprising that vocoded speech is significantly different to System E and F because both these systems do not contain any vocoded acoustic features whilst System C and D do. Listeners perceived System C and D to be equal in naturalness. This suggests that MCC and F0 features that are independent do not heavily influence naturalness perception. Also, listeners perceived System E and F to be equal in naturalness. This suggests that the predicted duration in the full DNN TTS did not alter listeners' naturalness perception.

In terms of self-reported cognitive load, human speech was rated as the easiest to listen to whilst speech produced by the full DNN TTS system was the hardest. Human speech was found to be significantly easier to listen to compared to all systems with the exception of vocoded speech. System C (vocoded MCC, predicted F0) and System D (vocoded F0, predicted MCC) were found to be easier to listen to compared to System F (full DNN TTS). It is evident that as we step from System A to System F, the cognitive load gradually increases which aligns with our predictions.

A significant negative correlation between the naturalness and cognitive load was found in Exp. 1 (corr=-0.55). In Experiment 2 in Chapter 4, a significant negative correlation between naturalness and cognitive load was also found. These findings suggest that when listening in quiet, listeners' tend to use their perception of naturalness as a measure of how cognitively demanding the speech is to listen to.

[3]Please note: All statistical results can be found in Appendix D.

Results: Growth Curve Analysis

Table 6.5: GCA parameter estimates of each time term and system in Exp. 1

System	Intercept	Linear	Quadratic	Cubic
A (Human)	1.78	16.68	-10.26	-9.11
B (Vocoded)	1.36	11.46	-9.64	-5.3
C (Predicted F0)	1.61	9.55	-11.08	-5.45
D (Predicted MCC)	0.34	8.98	-5.28	-8.42
E (Natural Duration)	1.82	9.14	-16.89	-7.87
F (full DNN)	1.01	11.96	-7.49	-7.57

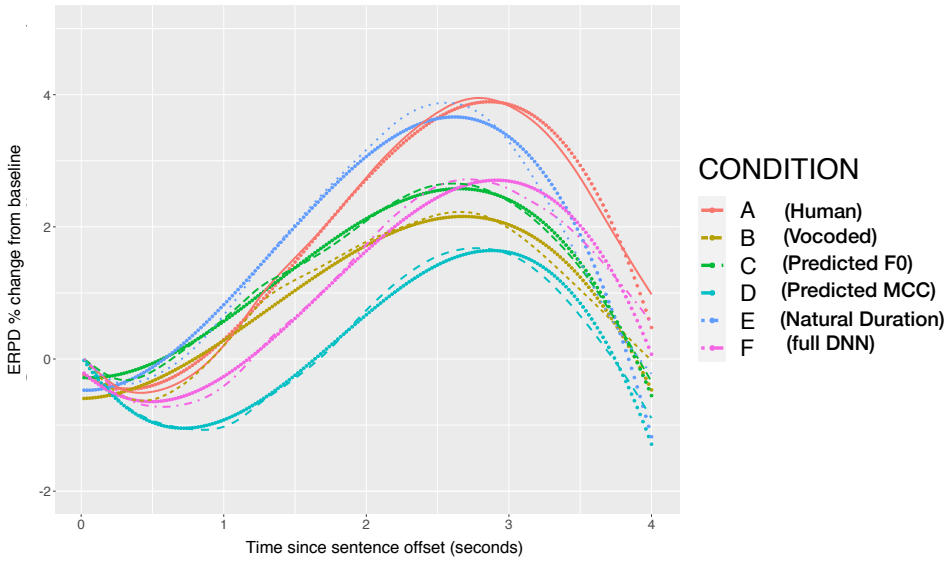


Figure 6.1: Time series line graph of the raw pupil response (dotted) and cubic model fit (solid) averaged across all participants for each system in Exp. 1 (Quiet).

Intercept The mean pupil response for all systems were found to be significantly[4] different from each other with the exception of System A and E. System D (Pred MCC & Voc F0) has the lowest mean pupil response and human speech and System E (Natural Duration) has the highest. Human speech, which is expected to have the lowest mean, evokes the highest mean pupil response. In Chapter 4, we observed the trend that high quality speech evoked a greater mean pupil response than low quality speech with the exception of human speech. However, we see that in these results, human speech evokes the greatest mean. As pointed out in Part I, it is more likely that when

[4]Please note: All statistical results can be found in Appendix D.

listening in quiet, the pupil response is indexing engaged attention and/or willingness to pay attention. The findings for this experiment show that listeners perceived human speech to be the easiest system to listen to, yet the mean pupil response is high. Thus, these findings further support the notion in Section 4.4.2, that the pupil response is indexing more the level of engagement rather than listening difficulty in quiet. Therefore, System E and human speech demand the greatest engaged attention whilst System D and F (full DNN TTS) demand the least. It is also evident, that the parameter estimate for the natural speech shown in Table 4.19 in Chapter 4 was low (2.09) compared to all TTS systems (> 3.5). The parameter estimates presented in this chapter in the intercept term fall in a range between 0.34 and 1.82 which suggest that in the case of speech produced by a DNN-speech synthesizers the pupil response appears to be indexing similar properties for both human and synthetic speech. This finding indicates the the demand of cognitive resources between human and synthetic speech are converging and therefore increases the likelihood that the pupil response is indexing similar properties in both cases.

Linear term Human speech has the steepest slope and significantly different to all other systems. Systems B and F were found to be equivalent and Systems C, D and E were found to be equivalent. Following on the notion of the pupil response indexing level of engagement, these findings suggest that human speech demands the highest level of engagement. In contrast, Systems C, D and E demand a lower level of engagement that human speech. Furthermore, the parameter estimates in the linear term in this experiment are all much lower than those presented in Chapter 4, Exp. 1 (except Low-Quality HMM). Apart from Low-Quality HMM in Chapter 4, Exp. 1, Natural had the smallest gradient. Therefore, this confirms that DNN-speech and human speech are converging in the amount of resources they demand.

Quadratic term System E has the sharpest peak and is significantly different to all other systems. System D has the flattest peak and is also significantly different to all other systems. System A, B and C were found to be equivalent. The full DNN TTS system has the second flattest peak. In experiment conducted in Section 4.4.2, human speech also evoked the flattest peak. Therefore, it appears that with high quality speech, flatter peaks are more favourable when listening in quiet than sharper peaks. However, System F contradicts this notion, it evokes a flatter peak than human speech. In comparison to results presented in Table 4.19 in Chapter 4, we observe human speech to have a parameter estimate of -6.33 whilst all other systems range between -31.53 and 42.02. The parameter estimates for the quadratic term in this experiment are all below -16.89. Therefore, although TTS has a lower parameter estimate than human speech in this experiment, overall it appears that all systems compared in this experiment elicit peaks that can be described as more flat than sharp.

Cubic term Human speech has a steep falling slope which is associated with high levels of engagement and is found equivalent to Systems D, E and F. Systems B and C have the least steepest slopes. In general, these findings relate to the findings discussed in all other time terms.

Summary

In this experiment, we investigated the cognitive load when listening to human speech and synthetic speech produced by a DNN-based speech synthesizer in quiet. Intelligibility across all systems compared were found equivalent. Therefore, any cognitive load differences observed will be a result of other properties of the speech signal beyond intelligibility. The self-reported measures indicate that human speech is the easiest to listen to - as was expected - whilst synthetic speech produced by the full DNN TTS system was the hardest to listen to. In Chapter 4, Hybrid, Unit Selection and HMM were all reported to have medians of 3 with respect to self-reported cognitive load. It is interesting that for the full DNN TTS system, which is reported as more natural sounding than Unit Selection and HMM, listeners still rated the cognitive load as a 3 (slightly difficult to listen to). Human speech was rated a 1 (easy to listen to) and vocoded speech was rated as 2, which suggests that the contributions are are not caused by the vocoder and are introduced by either the acoustic or duration model.

The pupil response showed contrasting findings to those reported in the self-reports where human speech evoked a pupil response with properties that are associated mostly with high listening effort (high mean, steep rising and falling slopes). However, these properties are *high* relative to the systems compared in this experiment. In comparison to the parameter estimates of the systems evaluated in Chapter 4, we observe that in this experiment the parameter estimates in general are lower. Therefore, although human speech in this experiment elicited the highest parameter estimates, those parameter estimates were still lower in comparison to speech of poorer quality. Therefore, one needs to be careful not to equate high parameter estimates to high listening effort but rather as a comparison between the systems evaluated in a given experiment. In other words, human speech evoked a pupil response with properties that translate to demanding more resources than all other systems compared in this experiment. Therefore, as concluded in Chapter 4, when listening in quiet, the pupil response is likely to be indexing more the level of engagement as it is unlikely that human speech is more difficult to listen to than all other systems.

In terms of understanding the contributions of each vocoded speech parameter, human speech demands more resources than vocoded speech according to all time terms. System C demands more resources than vocoded speech in all time terms except in the linear term. Systems B, C and E demand more resources than System D in all time terms except in the cubic term. System E demands more resources than System C and F in all time terms except the linear term. Human speech demands more resources than synthetic speech in all time terms. System C is more resource intensive (engaging) than vocoded speech, which suggests that predicted F0 demands more resources than vocoded speech whilst predicted MCC demands less resources than vocoded speech. Copying MCC and F0 which keeps dependencies between these features intact in System E results in an increased use of cognitive resources than System C and D. Using natural duration over predicted duration results in increased use of cognitive resources whilst predicted MCC is found to use less resources in quiet. Furthermore, in Chapter 4, a key finding was that the properties of the pupil

response when listening to human speech contrasted with that of synthetic speech. However, when listening to speech produced by a DNN-speech synthesizer we find natural and synthetic speech reflect similar properties. Therefore, the cognitive demand of synthetic speech appears to be approaching a similar demand to human speech.

6.3.2 Experiment 2: Noisy condition

In this experiment, three sub-experiments were conducted. Each of them measuring the listening effort of each speech configuration in Table 6.1 in the presence of speech-shaped noise at SNRs -1dB (Exp. 2A) -3dB (Exp. 2B) and -5dB (Exp. 2C) respectively. These SNRs were chosen specifically such that the cognitive load is increased whilst intelligibility remains close to ceiling. In accordance with the estimated psychometric function in Cooke et al. [2013] which related keyword scores to SNR for speech-shaped noise, the expected keyword correct percentages at -1dB, -3dB and -5dB are approximately 80% for -1dB 60% for -3dB and 40% for -5dB for natural speech. We used these percentages as a guideline when setting the WER threshold for each SNR in our analysis. In other words, only trials that had a WER below the threshold will be included in the analysis.

Pre-processing

Details pertaining to the pre-processing carried out for Exp. 2 are summarized in Table 6.6.

Table 6.6: Experiment analysis details of Exp. 2, including the total number of participants that were included in the analysis with the total number of recruited participants in brackets, the total number of trials remaining after the trial exclusion criteria was applied with its respective percentage shown in brackets and the mean recall accuracy percentage

Experiment	Participants	No. of trials (%)	Mean Recall Accuracy %
2A	18(18)	1379 (85)	96
2B	19(19)	1446 (85)	87
2C	17(19)	1387 (81)	76

Results: Intelligibility

Recall accuracy in all noise experiments were greater than 75%. Therefore, even under noisy conditions intelligibility remained close to ceiling. In the -1dB SNR condition, recall accuracy dropped by 3% compared to listening in quiet. Decreasing the SNR to -3dB resulted in a 9% drop in accuracy and a further 11% drop when listening in the -5dB SNR condition. Table 6.7 shows the WERs for each sub-experiment in terms of each system evaluated.

In quiet, Human speech has the lowest WER and the full DNN TTS system has the highest but no significant[5] differences were observed. In Exp. 2A, human, vocoded, System C and System D were equivalent with the lowest WERs whilst Systems E and F were equivalent with the highest. In

[5]Please note: All statistical results can be found in Appendix D.

Exp. 2B, a similar trend of results were observed. Human, vocoded and System C were equivalent with the lowest WER whilst Systems D, E and F were equivalent with the highest. In Exp. 2C, the trend of WERs were the same as in Exp 2B, Systems D, E and F were all equivalent with the highest WERs. As expected, as the SNR decreased, the WERs increased for all systems evaluated.

Table 6.7: WER percentage of systems for each sub-experiment in Exp. 2. For ease of comparison, we also include the results of Exp. 1

System	**WER %**			
	Exp. 1	**Exp. 2A**	**Exp. 2B**	**Exp. 2C**
A (Human)	2	5	9	20
B (Vocoded)	4	4	7	19
C (Predicted F0)	4	4	7	14
D (Predicted MCC)	2	8	15	28
E (Natural Duration)	3	11	18	32
F (full DNN)	5	12	20	34

Results: Self-reported measures

Table 6.8: Self-reported measures (Naturalness Score, Nat – higher is better) and (Cognitive Load, CL – lower is better)

System	**Exp. 1**		**Exp. 2A**		**Exp. 2B**		**Exp. 2C**	
	Nat	**CL**	**Nat**	**CL**	**Nat**	**CL**	**Nat**	**CL**
A (Human)	4	1	4	2	4	3	4	3
B (Vocoded)	4	2	4	3	3	3	4	4
C (Predicted F0)	3	2	4	3	4	3	3	3
D (Predicted MCC)	3	2	3	3	3	4	4	4
E (Natural Duration)	2.5	2.5	3	3	4	4	4	4
F (full DNN)	3	3	3	4	3	4	4	4

The medians of the self-reported measures for Exp. 1 and 2 are presented in Table 6.8. Across all three sub-experiments in noise and in quiet, human speech was rated the most natural sounding. Perception of human speech was not affected when listening in quiet or in noise. As the SNR decreased, CL for natural speech increased - as expected - but difficulty was perceived the same for -3dB and -5dB. For speech produced by the full DNN TTS system, listeners perceived the speech as slightly natural for all conditions except the most difficult condition, in which listeners rated the speech as more natural. We observe that for all systems except System C, the systems were rated as being equally or more natural in the harder SNR condition than the easier SNR condition. This shows that listeners lose their ability to make proper judgements with regards to how natural the speech sounds when listening in adverse noise conditions such as -5dB SNR.

In terms of the self-reported cognitive load, human speech was generally the easiest to listen to, even in the presence of noise whilst speech produced by a full DNN TTS system was difficult to

listen to in all the noise conditions. Vocoded speech only became difficult to listen to in the -5dB SNR condition. Systems D and E became difficult to listen to at the -3dB SNR condition. System C, like human speech remained only slightly difficult to listen to even in the most challenging SNR condition. These findings indicate that in the presence of noise, synthetic speech is harder to listen to than human speech. Perceptually, the main contributions to increased cognitive load appear to be a result of the quality of the predicted spectral features and duration which is indicated by the results obtained by Systems D and E respectively.

A weak negative correlation between the naturalness and cognitive load was found in Exp. 2A and Exp. 2B (corr=-0.31 and corr=0.39). In Exp. 2C, the correlation is not at all significant. This result suggests that when listening in easier SNR levels, listeners are still able to pay more attention to speech naturalness (like they can in quiet) and therefore naturalness becomes a contributing factor when scoring CL. When listening conditions become more difficult like in the case of listening in -5dB SNR, it becomes difficult for listeners to perceive how natural the speech sounds and thus it is less likely to be considered when scoring the CL. As a result, the correlation between naturalness and cognitive load becomes insignificant.

Results: Growth Curve Analysis

Table 6.9: GCA parameter estimates of each time term and system in Exp. 2A

System	**Intercept**	**Linear**	**Quadratic**	**Cubic**
A(Human)	1.11	11.26	-2.39	-4.21
B (Vocoded)	1.15	11.77	-3.78	-6.84
C (Predicted F0)	0.2	0.89	-6.76	-4.02
D (Predicted MCC)	0.47	10.27	-4.41	-7.63
E (Natural Duration)	1.63	15.31	-7.32	-6.85
F (full DNN)	2.42	25.92	-4.95	-7.69

Table 6.10: GCA parameter estimates of each time term and system in Exp. 2B

System	**Intercept**	**Linear**	**Quadratic**	**Cubic**
A (Human)	3.90	26.09	-10.74	-4.7
B (Vocoded)	3.95	24.92	-13.27	-6.0
C (Predicted F0)	3.17	22.49	-14.22	-9.61
D (Predicted MCC)	4.48	32.45	-6.54	-4.33
E (Natural Duration)	3.67	29.14	-14.12	-9.32
F (full DNN)	5.37	42.43	-10.67	-9.08

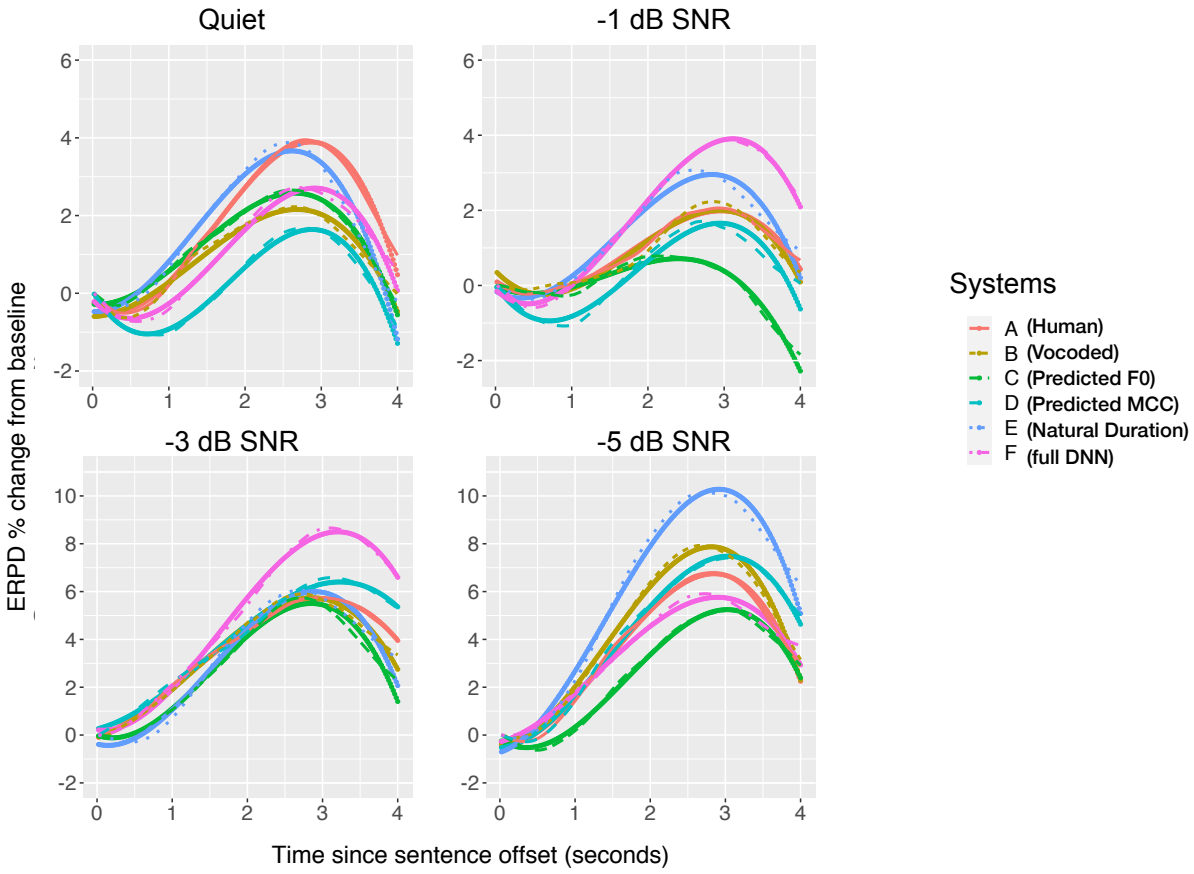


Figure 6.2: Time series line graph of the raw pupil response (dotted) and cubic model fit (solid) averaged across all participants, Exp.1 (top left), Exp 2A (top right), Exp. 2B (bottom left) and Exp.2C (bottom right)

Table 6.11: GCA parameter estimates of each time term and system in Exp. 2C

System	Intercept	Linear	Quadratic	Cubic
A (Human)	4.24	30.72	-18.64	-10.82
B (Vocoded)	4.97	34.46	-23.24	-13.47
C (Predicted F0)	2.62	24.85	-11.14	-10.23
D (Predicted MCC)	4.20	35.32	-13.40	-9.17
E (Natural Duration)	5.66	41.43	-23.34	-12.96
F (full DNN)	3.91	26.52	-14.50	-7.01

Intercept In all three noise experiments, System C (Voc MCC & Pred F0) has the lowest mean and is significantly[6] different to all other systems. In Exp. 2A and Exp. 2B, System F (full DNN TTS) evokes the greatest mean. In Exp. 2C, System F drops from the highest mean to the second lowest mean. We believe this could be due to the listener withdrawing effort/resources as a conse-

[6]Please note: All statistical results can be found in Appendix D.

quence of a task being too challenging to listen to which results in a smaller than expected pupil response.. Therefore, when listening to speech produced by a DNN-speech synthesizer, ceiling cognitive capacity is reached at the -3dB SNR condition. System A (human) and System B (vocoded) were found to be equivalent in Exp. 2A and 2B but in Exp. 2C System B is significantly different to all other systems. When listening in the easier SNR levels, vocoded speech appears to be processed in a similar manner to human speech but as the SNR levels become more challenging they begin to diverge. System D goes from the second lowest in Exp. 2A to the second highest in Exp. 2B and remains similar Exp. 2C.

Linear term In all three noise experiments, System C has the least steep slope and is significantly different to all other systems in Exp. 2A and Exp. 2B, but in Exp. 2C it is equivalent to System F (full DNN TTS). However, we know that System F reaches ceiling in -3dB and therefore a smaller response is expected. The full DNN TTS system has the steepest slope in Exp. 2A and Exp. 2B but drop in steepness in Exp. 2C. A similar observation for System F was observed in the intercept term, further supporting the notion that ceiling capacity for System F was reached in the -3dB condition. System A and B are found equivalent in Exp. 2A and Exp. 2B but diverge in Exp. 2C, like observed in the intercept term. This further supports the notion that human and vocoded speech are processed similarly in -1dB and -3dB SNRs. System D, as observed in the intercept, goes from the second lowest in Exp. 2A to the second highest in Exp. 2B and remains the same in Exp. 2C.

Quadratic term In Exp. 2A, Human speech has the flattest peak, System E has the sharpest peak. In Exp. 2B, System C and E, have the sharpest peaks. System D has the flattest peak. In Exp. 2C, System B and E are equivalent and have the sharpest peaks and System C has the flattest peak. Its strange how System C goes from the sharpest peak to the flattest peak between Exp. 2B and Exp. 2C. It is evident that as the SNR levels decreases and listening becomes more challenging, the pupil increases in sharpness across all systems except System C.

Cubic term System A and C were found to be equivalent in Exp. 2A and has the least steep falling slopes. System D and F have the steepest slopes. In Exp. 2B, System A and D have the least steepest slopes whilst Systems C, E, F have the steepest. System F (full DNN TTS) has a steeper slope to human speech in Exp. 2A and 2B but in Exp. 2C, System F has the least steepest slope. Again, suggesting that System F reaches ceiling at -3dB. Furthermore, this result suggests that human speech is still manageable to process even at the -5dB SNR. In Exp. 3C, System A, C and D have lower gradients than Systems B and E. All are slopes are steep.

To understand the individual contributions as to why DNN-based speech synthesis demands a higher cognitive load to human speech, we present in Figure 6.3 the cubic model fits of the average pupil responses for each system individually when listening in quiet and in each of the three SNR conditions. For Systems A and C, we observe that the mean pupil response is greater in quiet than

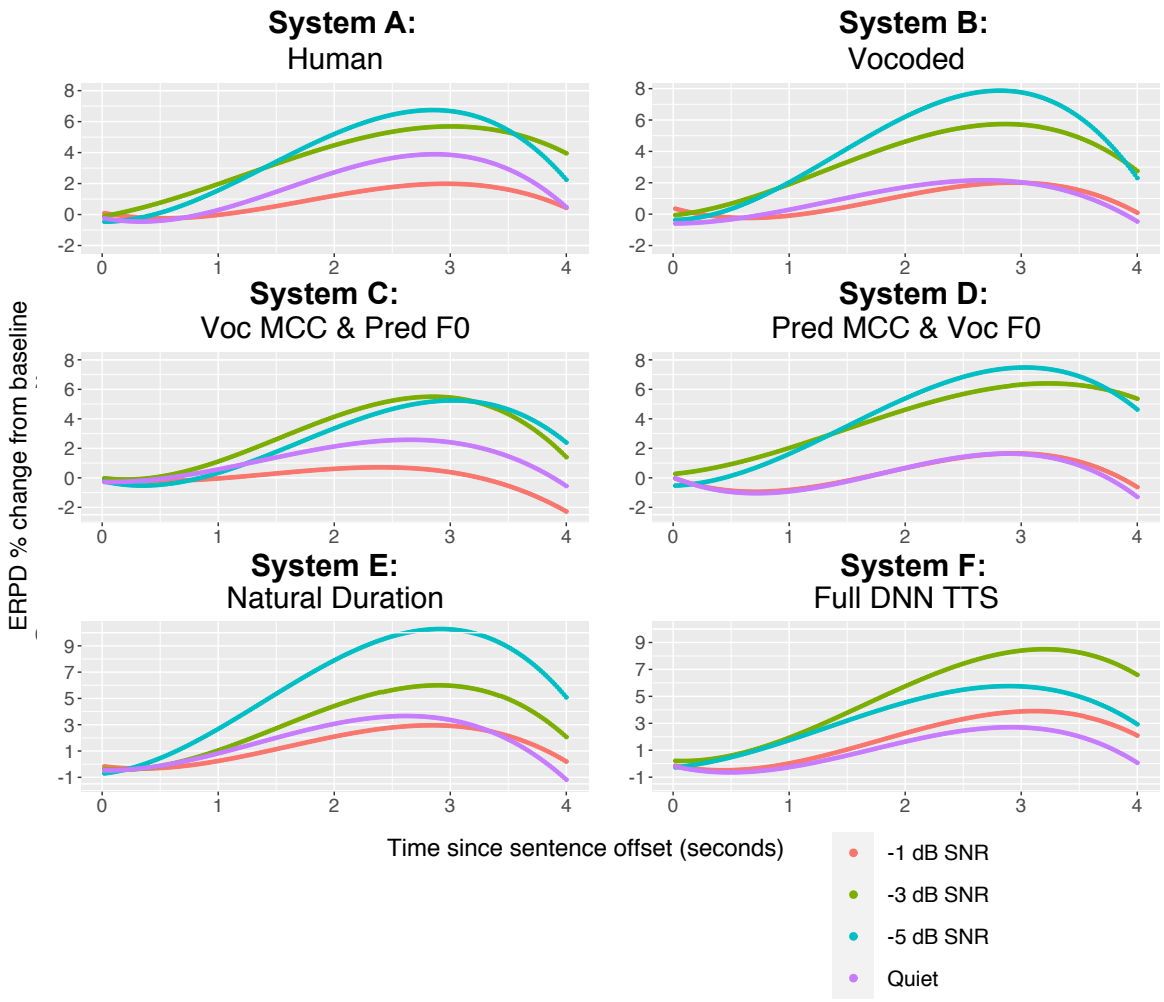


Figure 6.3: Time series line graph of cubic model fits for each system in Exp. 1 (Quiet) and Exp. 2 (Noise conditions)

in the -1dB noise condition. Systems B, D and E are equal. It is unlikely that listening in quiet is more challenging than listening in noise and therefore we believe that level of engagement is being indexed in the quiet condition. We also observe that for all systems, the mean pupil response increases as the SNR is decreased for all systems except for the System F (full DNN TTS). This finding was previously explained as a result of the listener withdrawing effort/resources as a consequence of a task being too challenging to listen to which results in a smaller than expected pupil response. Therefore, System F reaches ceiling capacity in -3dB which was confirmed in a number of time terms during the GCA analysis.

To understand the contribution to listening difficulty of each parameter we compare two systems at a time:

System A vs System B We observe that in quiet, System A evokes a greater mean pupil dilation than System B which implies System A is more resource intensive (engaging) than System B. In -1dB and -3dB SNR, the pupil response is equivalent. Therefore, in the presence of noise, it is likely that human speech and vocoded speech have the same listening difficulty. In -5dB SNR,

however, we observe that vocoded speech becomes more difficult to listen to than human speech.

System A vs System C In quiet, human speech is more resource intensive (engaging) than System C. In the presence of noise, in the easier SNR, human speech is less cognitively demanding than System C. However, in -3dB, human speech and System C converge whilst in -5dB, System C is equivalent to -3dB whereas human speech is a bit more cognitively demanding. This finding is abit more challenging to interpret without decreasing the SNR further. As it seems more likely that System C reaches cognitive capacity in the -5dB rather than being less demanding than human speech. This however cannot be confirmed unless we compare the result with an experiment with -7dB for example.

System B vs System C In quiet, System C evokes a greater mean pupil dilation than System B. Therefore, System C is considered to be more resource intensive (engaging) in quiet. The only difference between System B and C is that System C has predicted F0. This implies, that using predicted F0 features from an acoustic model results in speech that is more resource intensive (engaging) to listen to. In the presence of noise in all evaluated SNR conditions, vocoded speech and is slightly more difficult to listen to than System C. Therefore, predicted F0 has a positive influence on the difficulty of listening. It is difficult to know for certain whether System C is better than vocoded speech as there is uncertainty as to whether System C reaches ceiling capacity in -5dB.

System B vs System D System B and System D appear equally resource intensive (engaging). In the easier SNR, System B and System D are similar. In -3dB, they are also similar, however, the rising slope is a lot steeper for System D. In the harder SNR, System B is marginally more challenging than System D. Also, the increase between -3dB and -5dB for System D is smaller than System B which leads us to believe that System D is more challenging overall than System B. Therefore, predicted MCC seems to have a slightly different affect in how it is processed compared to System B.

System C vs System D System C is found to be more resource intensive (engaging) than System D. In the presence of noise, System D is more difficult to listen to than System C in all noise conditions. Therefore the acoustic model used in our DNN-speech synthesizers appears to model F0 features better than MCC features.

System E vs System B, C and D The results for these systems all follow a similar trend. System E evokes a greater pupil response than Systems B,C and D when listening in quiet. Therefore System E is more resource intensive (engaging) to listen to. System E evokes a greater response than Systems B, C and D in -1dB. In -3dB System B and E are equivalent and System D is greater than System E. In -5dB, System E evokes the greatest response compared to Systems B,C and D. In the easiest and hardest SNR System E is the most difficult to listen to. However, in the -3dB SNR System D is the most difficult to listen to. In general, this implies that features predicted by the acoustic model together with natural duration is more difficult to listen to than the other

configurations. The human cognitive processing system appears to be sensitive to the mismatch between natural duration and predicted features obtained from an acoustic model. A higher mean pupil response in the quiet condition implies that System E appears to be processed differently compared to any other system as it demands a high allocation of mental resources when listening in both quiet and noise.

System B vs System F System F evokes a higher pupil response than System B in quiet which shows that the full DNN TTS system is more resource intensive (engaging) to listen to than vocoded speech. It is interesting that System F is more resource intensive (engaging) than vocoded speech, this could be due to the System F having predicted F0. Since predicted F0 (as seen for System C) has a positive influence on the pupil response. However in the presence of noise, System F evokes a greater pupil response than System B except in the -5dB SNR condition. Therefore, listeners reach cognitive capacity when listening to System F at the -3dB SNR.

System E vs System F System E is more resource intensive (engaging) than System F. The only difference between System E and F is the duration. Therefore, speech with natural duration and consistent acoustic features are more resource intensive (engaging) to listen to than TTS which have consistent acoustic features but poor duration prediction. System F is more difficult to listen to than System E in all noise conditions except in -5dB SNR. As already mentioned, this is a result of listeners reaching cognitive capacity when listening to System F at the -3dB SNR as a consequence of fatigue.

6.4 Summary

The cognitive load of synthetic speech produced by a DNN-based speech synthesizer indexed by the evoked pupil response in noise was investigated in this experiment. Intelligibility across all system configurations were close to ceiling even at the most challenging SNR level. System E (Natural duration) and System F (full DNN TTS) were found to have the highest WERs whilst human speech has the lowest. In terms of the self-reported measures, moving from System A to F, naturalness deteriorated in -1dB. However, in the -3dB and -5dB SNR levels, naturalness scores converged which indicates that listeners found it difficult to make naturalness judgements under these adverse conditions. Similarly, for self-reported cognitive load, listeners rated System A as the easiest to listen to and System F as the most difficult. However, in -3dB SNR, System D, E and F were all equivalent and perceived to be the most difficult to listen to and in -5db SNR B, D, E and F were equivalent. Also, listeners rated some systems as being easier to listen to in -5dB than -3dB which we know is unlikely. These findings show that listeners struggle with self-reporting under adverse conditions which makes the validity of self-reported results under such conditions questionable.

The pupil results in quiet showed similar findings to those observed in Chapter 4. This confirms that listening in quiet is indeed not measuring listening difficulty but rather levels of engagement. In

noise, the pupil response results confirm that even high quality output generated by a DNN-speech synthesis system is still more difficult to listen to than human speech. In the more challenging SNRs such as -5dB, human speech appears to be more cognitively demanding than synthetic speech. Such a finding is unlikely and therefore a more plausible explanation is that listeners reach cognitive capacity when listening to synthetic speech in the -3dB SNR and therefore the pupil response becomes fatigued when listening in the -5dB condition. As a consequence, a smaller pupil response is observed. Results also showed that vocoded speech and human speech appear to be processed in a similar manner in the easy SNR noise conditions such as -1dB and -3dB but diverge in the more challenging SNR.

With respect to the individual contributions of each vocoder speech parameter, it is evident that the vocoder itself contributes to an increased cognitive load. We observe that, predicted MCC features in the acoustic model lead to an even greater demand of cognitive resources that exceeds the cognitive load already demanded by the vocoder. However, interestingly, predicted F0 appears to reduce the cognitive load demanded by the vocoder. However, we can not know this for certain. Nevertheless, this finding suggest that improved MCC prediction is important for reduced cognitive load whilst predicted F0 does not appear to contribute negatively to increased cognitive load. In addition, duration prediction is important and when the acoustic features and duration model are not working in harmony (as in System E) this appears to be severely detrimental to our cognitive processing system and demands a greater cognitive load. Although, the full DNN TTS system has consistency in terms of the duration and acoustic features, it still evoked the greatest load of all and therefore poor predicted duration and poor predicted MCC appears to be responsible for the increased cognitive load in this DNN-speech synthesizer.

In conclusion, listening to high quality TTS - such as that generated by a DNN driving a vocoder - still requires greater cognitive effort than human speech, under noisy conditions. The contributions of cognitive load in DNN-based speech synthesis are mainly due to poor MCC prediction and poor duration prediction. When combining speech parameters extracted from human speech with predicted acoustic features, dependencies between these features are destroyed. This alone appears to result in an increased cognitive load. These findings highlight 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 human speech. Sequence-to-sequence modelling addresses these limitations and is therefore investigated in the next chapter.

Chapter 7

Cognitive load of state-of-the-art speech synthesizers

In this chapter, we address our last research question, "Do modern TTS systems still demand greater cognitive load than human speech?". This is evaluated by measuring the cognitive load of current-state-of-the-art models using the pupillometry paradigm developed in Phase I. The work in this chapter is important for two reasons, 1) To determine whether such methods are still relevant enough to provide insightful information on current state-of-the-art models and 2) To determine whether these text-to-speech models still demand a higher cognitive load in comparison to listening to human speech.

We start this chapter by giving an overview of the state-of-the-art models that were evaluated in this chapter. Two experiments were carried out, the first experiment was conducted at the beginning of 2020 on the state-of-the-art models that existed at that time. However, due to the covid-19 pandemic my experiments were suspended. The second experiment was recently conducted to investigate the latest text-to-speech models that exist. In Experiment 1, Tacotron 2 and DC-TTS were the state-of-the-art models selected for Experiment 1. For Experiment 2, a newer implementation of Tacotron 2 and Fastspeech were selected. In addition, two other models were selected to investigate differences between the latest neural models and older models investigated in previous chapters. These two models include a commercial grade Unit Selection (Hybrid) system for comparison in the first experiment and a Feed-forward DNN model trained using Merlin for comparison in the second experiment. Pupillometry was the method undertaken, which remains the same as the method in Chapter 6. Therefore, we provide only a brief overview of the experimental set-up and dive straight into the results. Finally, we conclude this chapter by summarizing the key findings.

7.1 Introduction

In Chapter 6, we observed that some of the key contributing factors to increased cognitive load when listening to a DNN-based speech synthesizer in noise is due to poor mel-cepstral coefficient (MCC) prediction and duration prediction. We also found that when speech parameters that are

extracted from human speech are combined with acoustic features predicted from an acoustic model in terms of dependencies between these features are destroyed, thus this resulted in an increased cognitive load. This finding implies that modelling spectral features, F0 features and duration separately could be the reason for an increased cognitive load. Current state-of-the-art models address this problem as speech is now typically modelled in a single unified framework. Therefore, it will be important to know whether these new models are more difficult to listen to than human speech or are we moving in the direction where listening difficulty for synthetic speech is converging with the listening difficulty of human speech. Our hypotheses for the experiments in this chapter is that vocoded speech and the state-of-the-art speech synthesizers will demand similar resources as human speech whilst the other older non-neural based synthesizers will demand more resources.

7.2 State-of-the-art TTS

In Chapter 2, we introduced sequence-to-sequence-based speech synthesis and provided a high-level overview of two architectures, namely Tacotron 2 and DC-TTS. Both architectures comprise of an encoder-decoder architecture coupled with an attention mechanism. The main difference between them is that Tacotron 2 utilizes mostly RNN-based layers whilst DC-TTS utilizes convolutional layers. Additionally, Tactron 2 predicts mel-spectrograms which is passed directly to a neural vocoder which then generates the output speech waveform. In DC-TTS mel-spectrograms are also predicted but is then passed to an additional network, called spectrogram super-resolution network that converts the course mel-spectrograms to a full spectrogram which is passed to a neural vocoder for synthesis. Since newer models are introduced in this chapter, an overview of these models is provided here.

Tacotron 2 In Experiment 2, a newer version of Tacotron 2 was investigated. A pretrained model was used in this experiment taken from [Gölge, 2020]. This model is implemented in conjuction with the MultiBand-Melgan vocoder model which will be explained in the next section. The model was trained using a Double Decoder Consistency (DDC) for 130K steps using a single GPU. According to the developers of this codebase, in the previous version of Tacotron 2, the model suffered from attention alignment problems at inference time. These problems occur especially with long-text inputs or out-of-domain character sequences. This is where DDC was introduced as a method to fight against these alignment problem. The DDC method is based on two decoders working simultaneously with different reduction factors. One decoder (coarse) works with a large factor, and the other decoder (fine) works with a small reduction factor. DDC is designed to settle the trade-off between the attention alignment and the predicted frame quality tuned by the reduction factor. In standard models, larger reduction factors are used and therefore have more robust attention performance but due to over-smoothing the final acoustic features are coarser in comparison to using a low reduction factor. Therefore, DDC combines these two properties at training time as it uses the coarse decoder to guide the fine decoder to preserve the attention performance without a loss of precision in acoustic features. DDC achieves this by introducing an additional

loss function comparing the attention vectors of these two decoders. For each training step, both decoders compute their relative attention vectors and the outputs. Due to the differences in their respective reduction values, their attention vectors are different lengths. The coarse decoder produces a shorter vector compared to the fine decoder. In order to mitigate this, the coarse attention vector is interpolated to match the length of the fine attention vector. After forcing them into the same length we use a loss function to penalize the difference in the alignments. This loss is able to synchronize the two decoders with respect to their alignments.

Fastspeech 2 Fastspeech 2 is an improvement of Fastspeech [Ren et al., 2019]. Fastspeech is a novel text-to-speech system that comprises of a feed-forward network based on Transformers to generate mel-spectrogram in parallel for TTS. Unlike, Tacotron, it is non auto-regressive and therefore the inference process is much faster. In addition, it does not use the encoder-decoder based architecture like Tacotron 2 and DC-TTS. Attention alignments are extracted from an encoder-decoder based teacher model for phoneme duration prediction. This is then used by a length regulator to expand the source phoneme sequence to match the length of the target mel-spectrogram sequence for parallel mel-spectrogram generation. Experimental results showed that these types of models produce high quality speech like Tacotron models which are autoregressive. Fastspeech 2 attempts to speed up synthesis and inference time even more by directly training the model with ground-truth targets instead of using the simplified output from teacher model. Fastspeech 2 also introduces more variation information of speech such as pitch, energy and more accurate duration as conditional inputs. Duration, pitch and energy are extracted from the speech waveform and directly used as conditional inputs during training and inference. The Fastspeech 2 model used in this experiment was adapted from [Ren et al., 2020].

Multi-band MelGAN The MelGAN is a vocoder that is based on Generative Adversairal Networks (GANS). More specifically, it is a non-autoregressive feed-forward convolutional architecture that is capable of performing audio waveform generation. It is the first vocoder that uses a GAN setup for raw audio generation. Such a model replaces autoregressive models. Therefore, this model is substantially faster than other mel-spectrogram inversion alternatives without degrading speech quality [Kumar et al., 2019]. The generator is a fully convolutional feed-forward network with mel-spectrograms as input and raw waveforms as output. The achitecture is such that in the generator one can efficiently increase the induced receptive fields of each output time-step leading to better long range correlation. The discriminator follows a multi-scale architecture in which 3 discriminators have an identical network structure but operate on different audio scales. For the training objective, feature matching is used to train the generator [Larsen et al., 2016]. Multi-MelGAN is a later version of the model that improves the model in the following aspects: (1) it increases the receptive field of the generator, which is proven to be beneficial to speech generation. (2) the feature matching loss is substituted with the multi-resolution short time fourier transform (STFT) loss to better measure the difference between fake and real speech. Together with pre-training, this improvement was found to yield both better quality and better training stability. (3) the MelGAN is extended with

multiband processing where the generator takes mel-spectrograms as input and produces sub-band signals which are subsequently summed back to full-band signals as discriminator input. These improvements reduce the model complexity whilst still maintaining high quality output [Yang et al., 2021].

Figure 7.1 presents a table comparing the various component differences in the architecture of the various TTS systems evaluated in this thesis.

Component	Tacotron 2	DCTTS	Fastspeech 2	Merlin
Phonetisation	Text norm + character embedding	Text norm + (optional) phonetic dictionary	Text norm + character embedding	Text norm + phonetic dictionary
Contextual linguistic processing	Pre-net + Encoder	Text Encoder	Text Encoder	POS + linguistic positional features
Alignment	Location-sensitive attention with convolution and projection	Guided attention	Variance Adapter	HTS-based forced alignment
Duration prediction	Learned end-to-end and predicted using location-sensitive attention	Learned end-to-end and predicted using forcibly incremental attention	Duration Predictor	Separately-learned linguistic-features-only duration model
Acoustic predictor	2 LSTM layers + linear projection	Audio Decoder	Mel-spectrogram Decoder	Acoustic model
Acoustic features	Intermediate-res log-power mel-spectrogram + clipping	Low-resolution spectrogram converted to mel-spectrograms	Mel-spectrogram	Mel-spectrogram
Acoustic feature post-processing	Post-net	SSRN	Waveform Decoder	MLPG + postfilter
Signal generation + signal post-processing	WaveRNN	WaveRNN	Multi-band MelGAN	Multi-band MelGAN

Figure 7.1: Table comparing the various component differences in the architectures of the various TTS systems evaluated in this thesis.

7.3 Methodology and Implementation

Data The LJSpeech dataset Ito and Johnson [2017] consisting of speech sampled at 22.1 kHz from an American female speaker was used to train all systems evaluated in this chapter.

Model training The Tacotron 2 model in the first experiment was trained using open-source code from a repository on github Fatchord [2019]. In the second experiment, a newer pre-trained Tacotron 2 model was used using a different code-based Gölge [2020]. The DC-TTS model was trained using an in-house CSTR model of DC-TTS Watts [2019]. The Fastspeech 2 model was

trained using an in-house CSIR[1] model that is not publicly available. The Unit Selection (Hybrid) model is a commerical grade system that is also not publicly available. An adapted version of Merlin was used to train a feed-forward neural network as an example of an SPSS TTS system Govender [2019].

Audio samples Audio stimuli presented to the participants were sentences generated by the four speech synthesizers described above. Sentences generated by the respective vocoders (WaveRNN and Multi-band melgan) and the human speech used to train them were included for comparison. The test sentences for Experiment 1 and Experiment 2 differ. It was important to ensure that the test data was not used during training in any of the models. Similar to the previous experiments in this thesis, a baseline pupil size is required for analysis. Therefore, each audio stimulus had a buffer of at least two seconds immediately before the onset of the sentence in the audio. Since the experiments in this chapter were administered in noise, the buffer in this experiment consisted of 2 seconds of speech-shaped noise. Similar to previous experiments, a three seconds buffer was placed at the end of the audio to give the pupil sufficient time to return to the baseline. To ensure a fair comparison of all stimuli, loudness normalisation was subsequently applied to all audio samples using the standard root-mean-square algorithm.

Set-up The same pupillometry set-up as described in Section 4.2, is used for the pupil data collection in this chapter. In summary, the speech stimuli were played to listeners through headphones in a noise-and light-controlled room. Simultaneously, the pupil size was measured using an eye tracker.

Presentation Audio stimuli were presented in 5 blocks plus a practice block at the start of the experiment. Blocks were arranged using a 5 x 5 Latin square design to ensure all listeners, systems and sentences were equally balanced. The practice block comprised of 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 blocks. Each of the subsequent blocks had 20 trials. All sentences within each block were randomized except the first 5 sentences which were kept fixed across participants as they were discarded during the analysis. At the end of each block, self-reported cognitive load and naturalness scores were collected on 5-point rating scales (1 - Very Unnatural, Very Difficult; 5 - Very Natural, Very Easy).

Participants Participants were recruited from university students and staff. All participants were native English speakers with no self-reported hearing problems.

Pupil data processing, **Trial Exclusions**, **Baseline Correction** and **Post-processing** all followed the same procedures as described in Section 4.2 in this chapter.

Analysis Growth Curve Analysis (GCA) as described in Section 4.3.2 was used to analyse and

[1]CSIR is the Council for Scientific and Inductrial Reseach in South Africa which is the company I currently work for

interpret the pupil data in this chapter.

7.4 Experiments

This chapter comprises of two experiments that aim to (1) investigate the cognitive load (specifically, listening difficulty) of speech produced by state-of the art models in comparison to human speech, (2) to determine whether the methods developed in this thesis are still relevant to the latest state-of-the-art models. Each experiment investigates these aims in relation to the state-of-the-art models at the time that the experiments were conducted. Experiment 1 investigates the state-of-the-art models in 2020. Experiment 2 investigates the state-of-the-art models in 2022.

7.4.1 Experiment 1: 2020 state-of-the-art models

In this experiment, 2 state-of-the-art TTS models were evaluated, namely, Tacotron 2 and DC-TTS. Both models were trained in combination with the WaveRNN vocoder. Therefore, to determine whether the cognitive load contribution is due to the acoustic model and not the vocoder we have also included samples generated by the vocoder alone. In addition, we included a commercial grade Unit Selection system to compare the state-of-the-art models with an older and non-neural network based TTS model. The human speech used to train these models was also evaluated.

Two sub-experiments were conducted. Each of them measuring the listening effort of the various systems in the presence of speech-shaped noise at SNRs -3dB (Exp. 1A) and -5dB (Exp. 1B) respectively. These SNRs were chosen specifically such that the cognitive load is increased whilst intelligibility remains close to ceiling. In Chapter 6, we observed that intelligibility was still high in the -3dB condition for the high quality speech synthesizers. Therefore in this experiment, we did not measure cognitive load in the -1dB condition so that we were certain that a sufficient amount of cognitive load was achieved.

As mentioned previously, the expected keyword correct percentages at -3dB and -5dB SNR are approximately 60% for -3dB and 40% for -5dB for natural speech. We used these percentages as a guideline when setting the WER threshold for each SNR in our analysis. In other words, only trials that had a WER below a given threshold was included in the analysis.

Pre-processing

Details pertaining to the pre-processing carried out for Exp. 1 are summarized in Table 7.1.

Results: Intelligibility

Recall accuracy in the -3dB and -5dB conditions were inline with the expectation of 60% and 40% recall accuracy acheieved typically for human speech. In the -5dB SNR condition, recall accuracy dropped by 22% compared to listening in -3db SNR condition. Table 7.2 shows the WERs for each sub-experiment in terms of each system evaluated.

Table 7.1: Experiment analysis details of Exp. 1, including the total number of participants that were included in the analysis with the total number of recruited participants in brackets, the total number of trials remaining after the trial exclusion criteria was applied with its respective percentage shown in brackets and the mean recall accuracy percentage

Experiment	**Participants**	**No. of trials (%)**	**Mean Recall Accuracy %**
1A	23(23)	1134 (66)	64
1B	15(15)	672(60)	49

In Exp. 1A, WaveRNN (vocoded speech) has the lowest WER and is closely followed by human speech and no significant[2] differences between them were observed. DC-TTS had the highest WER and was significantly different to all other systems. WaveRNN and Human were also found to be significantly different to Unit Selection.

In Exp. 1B, we observe a similar trend, where human and WaveRNN had the lowest WERs and were not significantly different to one another. DC-TTS had the highest WER and was found to be significantly different to human and WaveRNN. As expected, as the SNR decreased, the WERs increased for all systems evaluated.

Table 7.2: WER percentage of systems for each sub-experiment in Exp. 1

System	**WER %**	
	Exp. 1A	**Exp. 1B**
Human	30	42
Tacotron 2	35	50
DC-TTS	51	65
WaveRNN	27	43
Unit Selection	40	54

Results: Self-reported measures

Table 7.3: Self-reported measures (Naturalness Score, Nat – higher is better) and (Cognitive Load, CL – lower is better)

System	**Exp. 1A**		**Exp. 1B**	
	Nat	**CL**	**Nat**	**CL**
Human	3	4	3	3
Tacotron 2	3	4	3	3
DC-TTS	2	5	2	5
WaveRNN	3	4	3	3
Unit Selection	3	4	2	4

The medians of the self-reported measures for Exp. 1 are presented in Table 7.3. In Exp.1A, participants found it challenging to seperate the naturalness across conditions. A general perception

[2]Please note: All statistical analysis results can be found in Appendix E

of slightly natural was rated for all systems except DC-TTS which was rated unnatural. In Exp.1B, we observe a similar trend, however in this noise condition, participants perceived Unit Selection to also be unnatural. These results correlate with previous suggestions made in earlier chapters that listeners tend to lose their ability to make proper judgements with regards to how natural the speech sounds when listening in adverse noise conditions such as -3dB and -5dB SNR.

In terms of the self-reported cognitive load in Exp. 1A, all systems were found to be difficult to listen to with DC-TTS being the most difficult. In Exp. 1B, human speech, WaveRNN and Tacotron 2 were perceived to be easier to listen to than DC-TTS and Unit Selection. These findings indicate that in the presence of noise, Tacotron 2 and WaveRNN are perceived to demand the same amount of cognitive load as human speech.

A weak negative correlation between the naturalness and cognitive load was found in Exp. 1A and Exp. 1B (corr=-0.31 and corr= -0.24). This result suggests that when listening conditions become more difficult like in the case of listening in -3dB and -5dB SNR, it becomes difficult for listeners to perceive how natural the speech sounds and thus it is less likely to be considered when scoring the CL. As a result, the correlation between naturalness and cognitive load is weak.

Results: Growth Curve Analysis

Table 7.4: GCA parameter estimates of each time term and system in Exp. 1A

System	Intercept	Linear	Quadratic	Cubic
Human	3.70	29.74	-7.42	-0.22
Tacotron 2	3.16	35.92	-11.41	-11.02
DC-TTS	4.66	31.33	-23.12	-0.81
WaveRNN	3.25	34.38	-1.03	-2.96
Unit Selection	4.31	40.81	-9.65	-7.10

Table 7.5: GCA parameter estimates of each time term and system in Exp. 1B

System	Intercept	Linear	Quadratic	Cubic
Human	5.80	42.11	-33.09	-13.42
Tacotron 2	6.32	39.53	-41.82	-17.49
DC-TTS	5.31	15.12	-41.91	-3.2
WaveRNN	5.65	41.23	-30.07	-15.99
Unit Selection	6.20	39.56	-34.22	-6.24

In terms of GCA, we expect to see estimates which reflect that human, vocoded, DC-TTS and Tacotron 2 would be demand an equivalent amount of cognitive resources whilst Unit Selection (Hybrid) would demand the most.

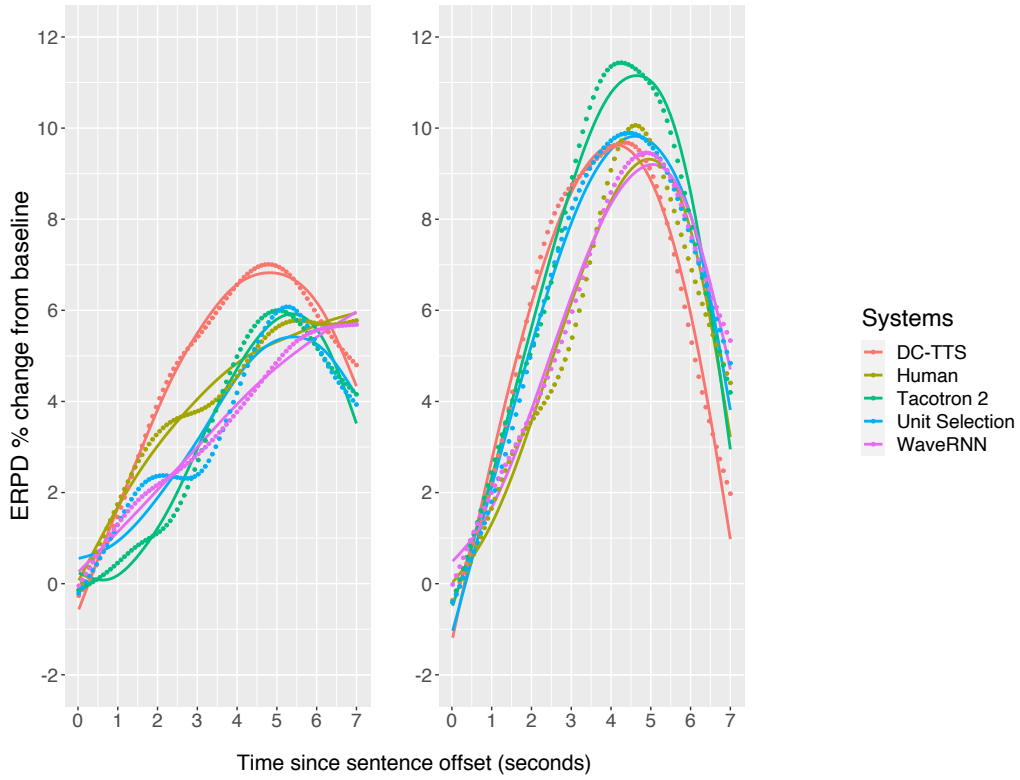


Figure 7.2: Time series line graph of the raw pupil response (dotted) and cubic model fit (solid) averaged across all participants, Exp.1A (left), Exp 1B (right)

Intercept For all systems, the intercept increases in value from the -3dB condition to the -5dB condition, which confirms that the cognitive load is greater in the -5dB condition than in the -3dB condition. In Exp. 1A, Tacotron 2 and WaveRNN have the lowest means and is significantly different to DC-TTS, Unit Selection and human speech. In other words, Tacotron 2 and WaveRNN are equivalent. DC-TTS has the largest mean and is significantly[3] different to all other systems. Unit Selection has the second largest mean and is also found to be significantly different to all other systems. Interestingly, even though human speech did not evoke the lowest mean, it is found to be significantly different to all other systems. In Exp. 1B, DC-TTS evokes the lowest mean. It is unlikely for DC-TTS to have the lowest cognitive load as it was perceived to be the most difficult and most unnatural condition which was observed in the subjective scores as well as in the -3dB condition. In previous chapters, we observed that when ceiling is reached, a reduced response is evoked. This finding therefore suggests that DC-TTS was too difficult to listen to in the -5dB condition and as a result a smaller response than expected is observed. DC-TTS was found to be signicantly different to all other conditions. WaveRNN has the second lowest mean and is significantly different to Tacoton 2, and Unit Selection. WaveRNN and human speech are thus equivalent. Tacotron 2

[3]Please note: All statistical analysis results can be found in Appendix E

goes from the lowest mean to the highest mean in -5dB and is found to be significantly different to human speech, WaveRNN and DC-TTS. Thus, Tacotron 2 and Unit Selection are equivalent. Across both noise conditions it is evident that human speech and vocoded speech demand the least cognitive load.

Linear term In the linear term, we observe that the slope is steeper or equivalent for all conditions except DC-TTS in the -5dB condition which can be explained by DC-TTS reaching ceiling in the -3dB condition, resulting in a less steep slope in the -5dB condition. In Exp. 1A, human speech evokes the least steepest slope and is found to be significantly different to all conditions except DC-TTS. DC-TTS is found to be significantly different to Tacotron 2 and WaveRNN. Unit Selection has the steepest slope and is significantly different to all other systems. In Exp. 1B, ignoring DC-TTS, Tacotron 2 has the least steepest slope and human speech has the steepest. However, all systems are found to be equivalent except DC-TTS which is significantly different to all other systems. Therefore, this finding suggests that all systems except DC-TTS appear to demand the same cognitive load in the -5dB condition.

Quadratic term In the quadratic term, we observe that all peaks become sharper in the -5dB condition than the -3dB condition which confirms that cognitive load is higher in the -5dB condition. In Exp. 1A, WaveRNN evokes the flattest peak whilst DC-TTS evokes the sharpest. This is inline with other findings in the other terms that consistently show that DC-TTS demands the most cognitive load. DC-TTS is also found to be significantly different to all other systems. In Exp. 1B, WaveRNN evokes the flattest peak but is only significantly different to Tacotron 2 and DC-TTS. DC-TTS evokes the sharpest peak and is significantly different to Tacotron 2. Therefore, across both conditions we observe very similar trends. As seen in the intercept term, human and vocoded speech appear to demand the lowest cognitive load and DC-TTS the highest.

Cubic term All slopes become more steeper in the -5dB condition except Unit Selection which appears similar. In Exp. 1A, human speech has the least steepest falling slope whilst Tacotron 2 has the steepest falling slope. Human speech is significantly different to all other systems except DC-TTS. DC-TTS has a smaller gadient than expected. WaveRNN is significantly different to Tacotron 2 and Unit Selection. Tacotron 2 is significantly different to Unit Selection. Therefore, all systems appear to behave differently in this time term. In Exp. 1B, DC-TTS has the least steepest falling slope. This is in line with all other terms that suggest that DC-TTS evokes a smaller response due to the listener possibly withdrawing resources as a result of the task being too challenging. Tacotron 2 again has the steepest falling slope. Tacotron 2 is significantly different to human speech and Unit Selection. Unit selection is significantly different to human speech. Once again most conditions behave differently.

In Figure 7.3 the cubic model fits for each system's evoked ERPD when listening in each of the SNR conditions is presented individually. For all systems, we observe that the pupil response

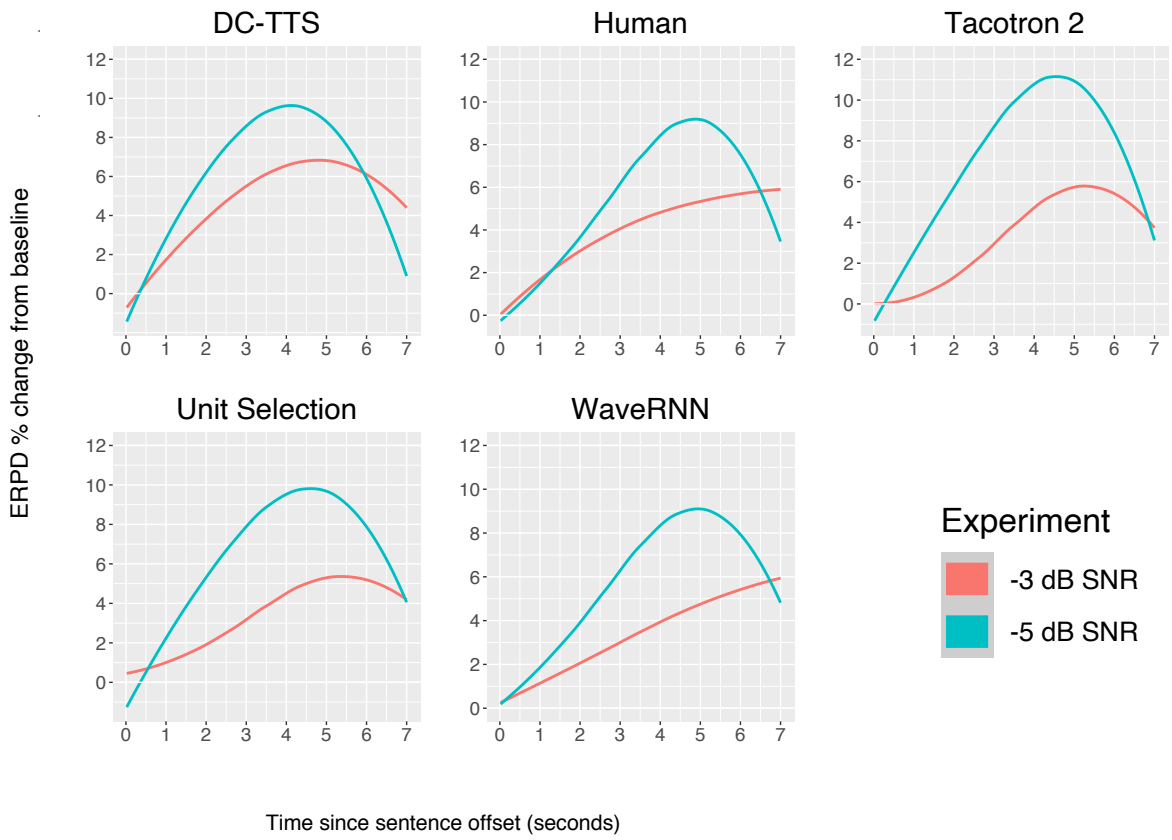


Figure 7.3: Time series line graph of cubic model fits for each system in Exp. 1A (-3dB SNR) and Exp. 1B (-5db SNR)

is greater in the -5dB noise condition than the -3dB noise condition. This finding confirms that listening difficulty is being measured consistently across all conditions.

We also observe that in the -3dB condition, human and vocoded (WaveRNN) speech behave similar whilst Tacotron and Unit Selection are similar. In the -5dB condition, WaveRNN and human speech behave similarly once again whilst Tacotron 2 appears to be the most difficult to listen to. Unit Selection and DC-TTS behave similarly but based on the findings in the GCA analysis, we know that DC-TTS evokes a smaller pupil response than expected. This corresponds with the self-reported measures where DC-TTS has the highest WER, is rated the most unnatural and is perceived to be the most difficult to listen to. In Figure 7.3, we observe that the increase from the -3dB condition to the -5dB condition for DC-TTS is the smallest compared to any of the other systems, showing further evidence that a smaller pupil response is being evoked.

Summary

The cognitive load of synthetic speech produced by state-of-the-art high quality text-to-speech systems were investigated in this chapter. This was performed by measuring the evoked pupil response when listening to speech produced by these systems in the presence of noise, specifically -3dB SNR and -5db SNR.

All systems scored below the expected WER thresholds for the -3dB (40%) and -5dB (60%) condition except for DC-TTS. DC-TTS exceeded the threshold for both conditions. Based on this finding, DC-TTS was the poorest performing system from all the systems compared.

For the self-reported measures, we observed that all systems were perceived equally natural in the -3dB condition except for DC-TTS which was rated unnatural and for self-reported cognitive load, all conditions were difficult to listen to except DC-TTS which was found to be the most difficult to listen to. Similarly, in the -5dB condition, most systems were rated equally for both naturalness and self-reported cognitive load except DC-TTS as well as Unit Selection. Whilst Unit Selection was rated the same naturalness as DC-TTS, it was however perceived to be slightly less difficult to listen to than DC-TTS.

The pupil response results showed that in the -3dB condition, human speech for most time terms corresponded with demanding the least cognitive load whilst vocoded speech closely followed. This finding suggests that vocoded speech produced by neural vocoders such as WaveRNN demand similar cognitive load as human speech when listening in noise. Tacotron 2 in two terms was found to be equivalent to vocoded speech. However, in other terms it showed that it demands more cognitive load than human and vocoded. Unit Selection demanded the second highest cognitive load across most terms and DC-TTS demanded the highest. It is interesting that the Unit Selection system performs better than DC-TTS as we expected a neural architecture to demand less cognitive resources than a non-neural based architecture. It is important to note that this specific Unit Selection system is a commercial grade system and whilst it does not directly use neural networks in its architecture, neural networks are used to assist with missing units in the system. In other words, a neural network is used to produce speech units that are missing in the database. Thereby making it Hybrid. It differs to the Hybrid system used in Chapter 4, as this one uses DNNs and not HMMs.

In the -5dB condition, the systems become more difficult to tease apart. Human speech demands the least listening effort and vocoded speech perform similarly in most time terms as observed in the -3dB condition. Tacotron 2 demands the most listening in -5dB. Unit Selection perform similarly and DC-TTS. As previously explained it appears that DC-TTS reaches ceiling capacity in the -3dB condition and as a consequence the pupil response evoked is smaller than expected. Unit Selection also has low estimates compared to the other conditions for all time terms except the mean. Therefore, it is possible that Unit Selection also reaches cognitive capacity in the -3dB condition but it isn't as obvious to see as DC-TTS.

Overall, these results show that vocoded speech which is becoming indistinguishable from human speech demands a similar amount of cognitive load as human speech. Systems such as Tacotron 2, which was one of the state-of-the-art models (at the time the experiment was conducted), demands slightly more load than vocoded and human speech. DC-TTS, despite being a neural text-to-speech

system performed the worst and closely followed was Unit Selection which was expected to demand more listening effort than the fully neural TTS systems.

In conclusion, speech produced by state-of-the-art TTS is moving in the direction of demanding a similar amount of cognitive load as human speech. Results in this chapter show that vocoded speech produced by a neural vocoder demands a similar amount of cognitive load as human speech. Therefore, the contribution of an increased cognitive load in Tacotron 2 and DC-TTS comes from the acoustic model. Tacotron 2 and DC-TTS are both sequence-to-sequence-based models yet performed so differently in these experiments. As shown in Figure 7.1, one key difference between these two systems is that Tacotron 2 predicts mel-specotrograms directly whereas DC-TTS first predicts spectrograms and then converts them to mel-spectrograms that are then passed on to the vocoder. This could be a differentiating factor in the speech signal that contributes to an increased cognitive load and explains why DC-TTS performed the worst. It will be interesting to investigate this further in a similar manner to the way that experiments in Chapter 6 were conducted. However, architectures like this are a lot more difficult to tease apart and therefore was not possible to do within the scope of this thesis.

7.4.2 Experiment 2: 2022 state-of-the-art models

In this experiment, 2 state-of-the-art TTS models were evaluated, namely, Tacotron 2 and Fastspeech 2. Both models were trained in combination with the Multi-band melGAN vocoder. Tacotron 2 in this experiment is different to the model used in the Experiment 1. This model is a newer version of Tacotron 2 that currently exists. Similar to Experiment 1, to determine whether the cognitive load contribution is due to the acoustic model and not the vocoder we have also included samples generated by the vocoder alone. In addition, we included an updated version of the Merlin text-to-speech system that was evaluated in Chapter 6. Similar to Experiment 1, this model was included to compare the state-of-the-art models with an older neural network based TTS model. The human speech used to train these models was also evaluated.

Three sub-experiments were conducted. Each of them measuring the listening difficulty of the various systems in the presence of speech-shaped noise at SNRs -1dB (Exp. 2A), -3dB (Exp. 2B) and -5dB (Exp. 2C) respectively. The -1dB SNR condition was included here due to the Merlin system previously observed to reach ceiling capacity in the -3dB SNR condition in Chapter 6.

As mentioned previously, the expected keyword correct percentages at -1dB, -3dB and -5dB are approximately 80%, 60% and 40% for natural speech respectively. We used these percentages as a guideline when setting the WER threshold for each SNR in our analysis. In other words, only trials that had a WER below a given threshold was included in the analysis.

Pre-processing

Details pertaining to the pre-processing carried out for Exp. 2 are summarized in Table 7.6.

Table 7.6: Experiment analysis details of Exp. 2, including the total number of participants that were included in the analysis with the total number of recruited participants in brackets, the total number of trials remaining after the trial exclusion criteria was applied with its respective percentage shown in brackets and the mean recall accuracy percentage

Experiment	Participants	No. of trials (%)	Mean Recall Accuracy %
2A	15(14)	695 (62)	75
2B	20(20)	910(61)	64
2C	25(24)	1112(59)	51

Results: Intelligibility

Recall accuracy in the -3dB and -5dB conditions were inline with the expectation of 60% and 40% recall accuracy acheieved typically for human speech. However, for the -1dB condition, recall accuracy was lower than expected due to the Merlin system performing poorly. In the -3dB condition, recall accuracy dropped by 10% from the -1dB condition and in the -5dB condition, recall accuracy dropped by 14% compared to the -3db SNR condition. As expected, as the SNR decreased, the WERs increased for all systems evaluated. Table 7.7 shows the WERs for each sub-experiment in terms of each system evaluated. In Exp. 2A, human speech had the lowest WER and was

found to be equivalent all other systems except Merlin. Merlin had the highest WER and was significantly[4] different to all systems. In Exp. 2B, Tacotron 2 had the lowest WER and was found to be significantly different to Merlin and Fastspeech 2. Similar to Exp. 2A, Tacotron 2, human and Multi-band melGAN are found to be equivalent. Merlin has the highest WER and was significantly different to all systems except Fastspeech 2. In Exp. 2C, the same trend continues where all systems are found to be equivalent except Fastspeech 2 and Merlin. Merlin again having the highest WER.

Table 7.7: WER percentage of systems for each sub-experiment in Exp. 1

System	**WER %**		
	Exp. 2A	**Exp. 2B**	**Exp. 2C**
Human	20	32	45
Tacotron 2	21	30	44
Fastspeech 2	27	39	53
Multi-band melGAN	22	33	48
Merlin	34	43	54

Results: Self-reported measures

Table 7.8: Self-reported measures (Naturalness Score, Nat – higher is better) and (Cognitive Load, CL – lower is better)

System	**Exp. 2A**		**Exp. 2B**		**Exp. 2C**	
	Nat	**CL**	**Nat**	**CL**	**Nat**	**CL**
Human	4	3	3	3	4	4
Tacotron 2	3	3	3	3	4	3
Fastspeech 2	3	4	3	4	3	4
Multi-band melGAN	3	3	3.5	2.5	4	3
Merlin	2	4	3	4	3	4

The medians of the self-reported measures for Exp. 2 are presented in Table 7.8. In Exp.2A, participants found human speech to be the most natural and Merlin to be the least natural. Human speech was significantly different to Fastspeech 2 and Merlin. Although Fastspeech 2 has a median of 3, the range of values fell between 2 and 3. Fastspeech 2 was found to be significantly different to Merlin. In Exp. 2B, participants found it challenging to separate the naturalness across systems. All systems had a median of 3 except Multi-band melGAN which has a slightly better rating for naturalness. Merlin, however, was found to be significantly different to all systems as the ratings for Merlin fell in the range of 2 and 3. In Exp. 2C, it is interesting that participants rated the systems higher than in the other two noise conditions. We observed the same finding in previous chapters. Human, Tacotron 2 and Multi-band melGAN were all rated the most natural whilst Fastspeech 2 and Merlin were rated slightly natural. Human speech was found to be significantly different to Fastspeech 2 and Merlin and Merlin was significantly different to all systems except Fastspeech 2.

[4]Please note: All statistical analysis results can be found in Appendix E

For the self-reported cognitive load in Exp. 2A, Merlin and Fastspeech 2 are perceived to be the hardest to listen to. Merlin was found to be significantly different to all systems except Fastspeech 2. All other systems were equivalent. In Exp. 2B, once again, Merlin and Fastspeech 2 were the hardest to listen to whilst Multi-band melGAN was the easiest to listen to. However, Multi-band melGAN was only found to be significantly different to Fastspeech 2 and Merlin. In Exp. 2C, Tacotron 2 and Multi-band melGAN are perceived to be easier to listen to compared to all other systems. It is interesting that human speech was not perceived as the easiest to listen to. Fastspeech 2 is found to be significantly different to Tacotron 2 and Multi-band melGAN. Human speech is found to be significantly different to Tacotron 2.

A weak negative correlation between the naturalness and cognitive load was found in all experiments (corr=-0.52, corr= -0.18, corr=-0.17 respectively). The correlation is highest in the -1dB condition but as the SNR decreases this correlation becomes weaker as observed in all previous experiments.

Results: Growth Curve Analysis

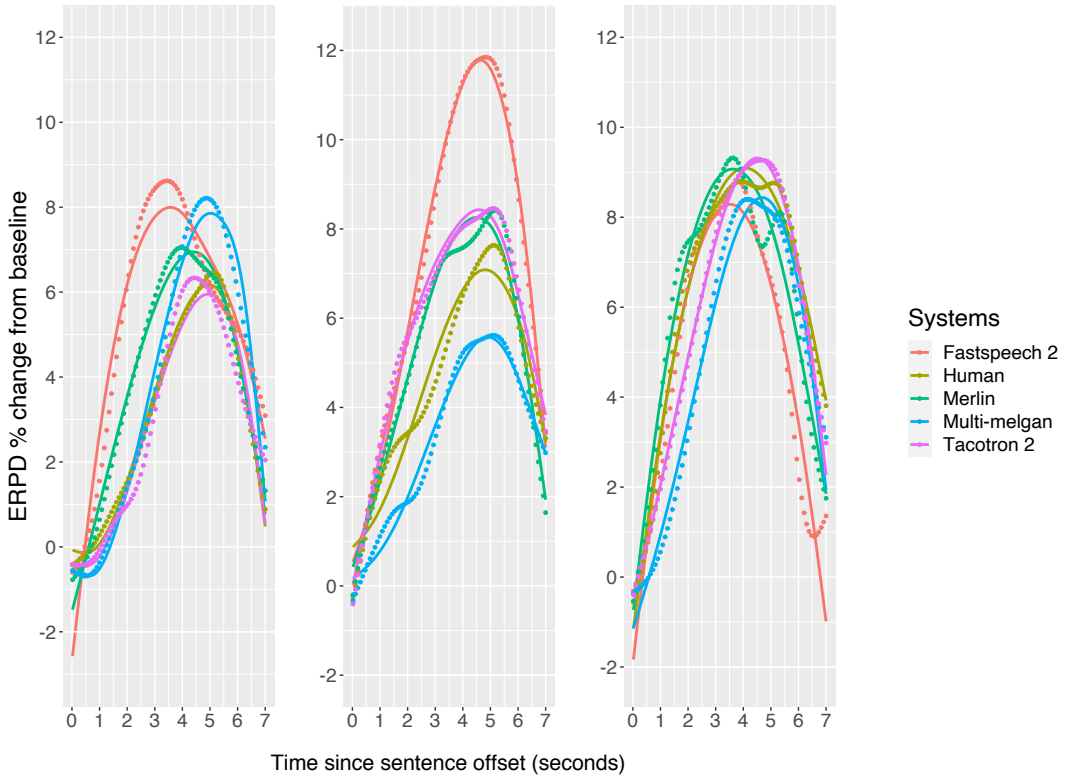


Figure 7.4: Time series line graph of the raw pupil response (dotted) and cubic model fit (solid) averaged across all participants, Exp.2A (left), Exp 2B (middle), Exp.2C (right)

Table 7.9: GCA parameter estimates of each time term and system in Exp. 2A

System	Intercept	Linear	Quadratic	Cubic
Human	3.61	38.13	-26.9	-21.24
Multi-band melGAN	4.69	50.68	-33.07	-26.33
Tacotron 2	3.53	39.21	-27.67	-19.84
Fastspeech 2	7.27	46.25	-47.64	-1.24
Merlin	5.57	39.59	-46.54	-13.70

Table 7.10: GCA parameter estimates of each time term and system in Exp. 2B

System	Intercept	Linear	Quadratic	Cubic
Human	4.63	25.38	-23.97	-12.61
Multi-band melGAN	4.55	41.64	-19.29	-12.02
Tacotron 2	6.38	37.63	-28.52	-5.05
Fastspeech 2	8.63	51.47	-49.56	-17.39
Merlin	6.00	29.2	-39.81	-14.29

Table 7.11: GCA parameter estimates of each time term and system in Exp. 2C

System	Intercept	Linear	Quadratic	Cubic
Human	5.54	23.23	-36.43	-1.59
Multi-band melGAN	5.46	37.63	-38.12	-10.00
Tacotron 2	6.24	37.29	-41.47	-12.57
Fastspeech 2	5.83	18.91	-52.12	-2.30
Merlin	5.73	16.01	-40.59	-1.67

Similar to Experiment 1, our predictions are that human speech, vocoded, Fastspeech 2 and Tacotron 2 will have parameter estimates that are equivalent and demand the least cognitive resources compared to Merlin which we believe will demand the most cognitive resources.

Intercept For all systems, the intercept increases in value or remains the same as the SNR decreases except for Fastspeech 2 and Merlin in Exp. 2C. This confirms that listening difficulty increases as the SNR decreases. Fastspeech 2 and Merlin have lower intercepts in the -5dB condition. This finding suggests that Fastspeech 2 and Merlin reached ceiling capacity in the -3dB condition and therefore a smaller mean pupil response was observed. The mean for Multi-band melGAN remained the same for the -1dB and -3dB conditions. Tacotron 2 remained the same for the -3dB and -5dB condition.

In Exp. 2A, Tacotron 2 had the lowest means and is significantly[5] different all systems except human speech. In other words, Tacotron 2 and human speech are equivalent. Fastspeech 2 has

[5]Please note: All statistical analysis results can be found in Appendix E

the largest mean and is significantly different to all other systems. Merlin and Multi-band melGAN were also found to be significantly different to all other systems. In Exp. 2B, Multi-band melGAN has the lowest mean but is found to be equivalent to human speech. Fastspeech 2 has the highest mean and is significantly different to all other systems. Tacotron 2 and Merlin were also significantly different to all other systems. In Exp. 2C, Multi-band melGAN has the lowest mean but is found to be equivalent to human speech. Tacotron 2 has the highest mean and is found to be significantly significant to all other systems. Merlin and Fastspeech 2 were also significantly different to all other systems. Merlin and Fastspeech 2 have lower means in -5dB than -3dB which suggests that ceiling was reached in the -3dB SNR condition.

Overall, human speech and vocoded speech behave similarly across all experiments and demand the lowest cognitive load. Fastspeech 2 demands the most cognitive load with Merlin closely following. Tacotron 2 goes from the lowest mean in -1dB to the highest mean in -5dB. This was also observed in Experiment 1.

Linear term In the linear term, we observe that the slope is less steeper in the -3dB condition than -1dB for all conditions except Fastspeech 2. Therefore, it appears that Fastspeech 2 is processed differently compared to any of the other systems. In the -5dB condition we observe a similar trend where all slopes are less steep compared to the slopes in -3dB except for Tacotron which remains the same. Fastspeech 2 and Merlin are much less steeper compared to any of the other systems. This is interesting, in all other experiments we observed that as the SNR decreases the slopes become steeper which does not seem to be the case for these experiments as we observe the reverse. Fastspeech 2 is the only condition observed to have a steeper slope and is the only system that has consistently shown that it demands the most cognitive load. In the Exp. 2A, human speech has the least steepest slope and is found to be significantly different to all conditions except Tacotron 2 and Merlin. Multi-band melGAN has the steepest slope and is significantly[6] different to all other systems. In Exp. 2B, human speech has the least steepest slope and is significantly different to all other systems. Fastspeech 2 has the steepest slope and is significantly different to all other systems. In Exp. 2C, Fastspeech 2 has the least steepest slope and is found to be equivalent to Merlin. From other findings we know that Fastspeech 2 demands the most cognitive load and therefore by having the least steepest slope suggests once again that it is likely that Fastspeech 2 reaches ceiling capacity in the -3dB condition. It might also be likely that Merlin too reaches ceiling capacity in the -3dB condition as observed in Chapter 6. Other than this, human speech has the least steepest slope whilst Multi-band melGAN and Tacotron 2 have the steepest slopes and are found to be equivalent.

Quadratic term In the quadratic term, we observe that the peaks are all flatter or similar in -3dB than -1dB except for Merlin. In the -5dB condition, all peaks become sharper as the SNR decreases. This leads us to believe that perhaps for all systems except Merlin, the -1dB condition is not measuring listening difficulty as an insufficient load is being placed on the listener as we

[6]Please note: All statistical analysis results can be found in Appendix E

observed in the quiet condition in previous chapters. In Exp. 2A, Human speech has the flattest peak and is found to be equivalent to Tacotron 2. Fastspeech 2 has the sharpest peak and is found to be equivalent to Merlin. In Exp. 2B, Multi-band melGAN has the flattest peak and found to be significantly different to all other systems. Closely following is human speech which is also significantly different to all other systems. Fastspeech 2 has the sharpest peak and is significantly different to all other systems. In Exp. 2C, human speech has the flattest peak and is found to be equivalent to Multi-band melGAN. Fastspeech 2 has the sharpest peak is significantly different from all other systems. We consistently find that human and vocoded speech demand the least amount of cognitive load whilst Fastspeech 2 demands the most.

Cubic term In the cubic term, all slopes are less steeper in the -3dB condition than the -1dB condition except for Fastspeech 2 and Merlin which was observed in the linear term as well. This once again leads us to believe that for the other systems, listening difficulty may not be indexed in the -1dB SNR. In the -5dB condition, all systems have less steeper slopes in the -5dB SNR than -3dB SNR except for Tacotron 2. This is very strange as no other term suggests that any of the other systems apart from Fastspeech 2 and Merlin reach ceiling capacity in the -3dB SNR. In Exp. 2A, Fastpeech 2 has the least steepest falling slope and is significantly[7] different to all other systems. Human, Tacotron 2 and Multi-band melGAN have the steepest falling slopes and are equivalent. In Experiment 1, the worst performing system has the least steepest slope and in this experiment we observe the same. Human speech, Multi-band melGAN and Tacotron 2 are found to be equivalent. In Exp. 2B, Tacotron 2 has the least steepest falling slope and is significantly different to all other systems. Fastspeech 2 has the steepest falling slope and is equivalent to Merlin. Human speech and Multi-band melGAN are equivalent. Fastspeech 2 goes from the least steepest in Exp. 2A to the most steepest in Exp. 2B. In Exp. 2C, human speech and Merlin are found equivalent and have the least steepest slope. Tacotron 2 and Multi-band melGAN have the steepest and are found to be equivalent.

In Figure 7.5, the cubic model fits for each system's evoked ERPD when listening in each of the SNR conditions is presented individually. For all systems except Fastspeech 2, we observe that the pupil response is greater in the -5dB noise condition than the -3dB noise condition which is consistent with what we expect when decreasing the SNR. This is confirmed in all time terms, where the values either significantly increase or remain similar between the -3dB condition and -5dB condition. This finding confirms that listening difficulty is being measured consistently across all conditions in the -3dB condition and -5dB condition. Fastspeeech 2 is the only condition where the ERPD for -3dB is significantly greater than the -5dB condition. In the growth curve analysis, we also saw decreasing values for Fastspeech 2 which opposes our expectation that values should increase when we decrease the SNR. Thus we are lead to believe that ceiling capacity is reached in the -3dB condition for Fastspeech 2.

For the -1dB condition, we observe a different trend to the one we expect. The ERPD for the

[7]Please note: All statistical analysis results can be found in Appendix E

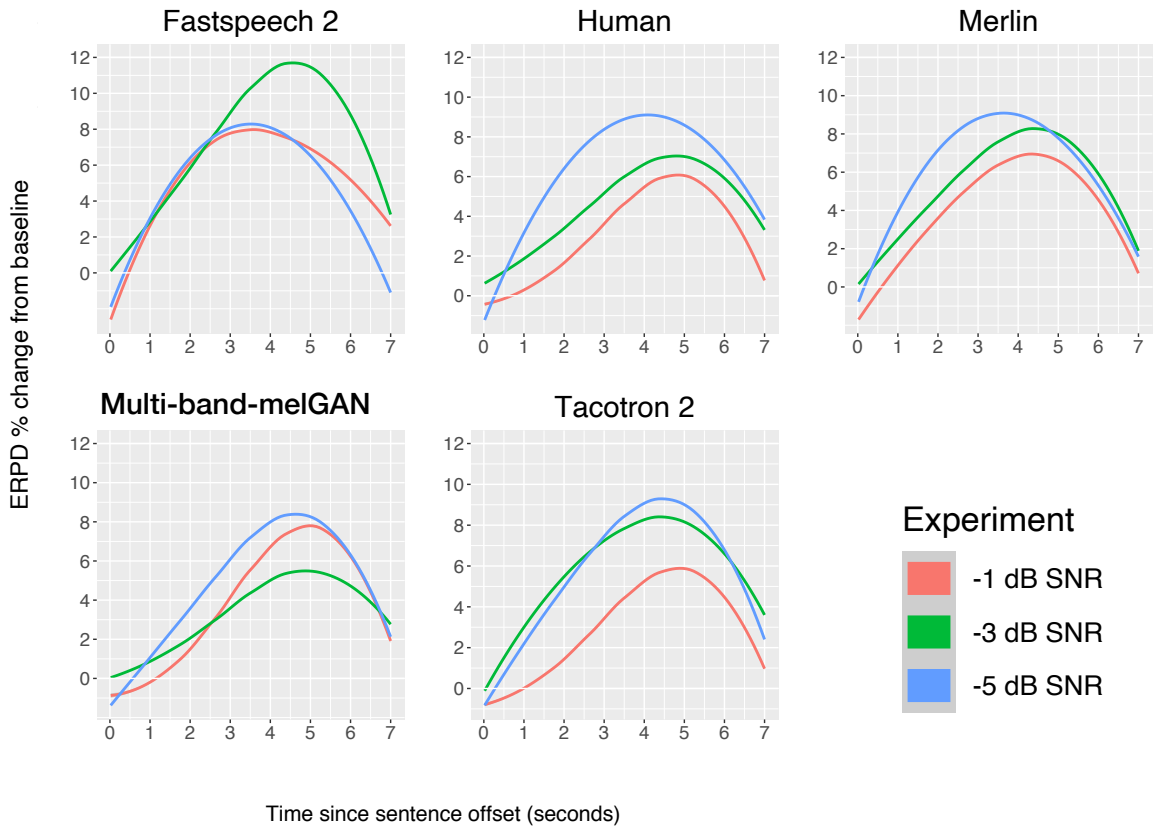


Figure 7.5: Time series line graph of cubic model fits for each system in Exp. 2A (-1dB SNR), Exp. 2B (-3dB SNR) and Exp. 2C (-5dB SNR)

-1dB condition appears to be lower than the -3db condition only for human speech, Merlin and Tacotron 2. This observation leads us to believe that listening difficulty may not be consistently measured across all of the conditions evaluated. In previous chapters, we observed that when insufficient load was placed on the listener, the pupil response appeared to index levels of engagement rather than listening difficulty. Given that most of the systems in this experiment is of high quality, it is likely that the -1dB condition is also not exerting a sufficient load and as a result the pupil response may not be indexing listening difficulty alone. In Chapter 6, we could easily identify that the pupil response was indexing something other than listening effort in quiet but upon closer inspection, we find that the ERPD response of vocoded and human speech in Chapter 6 was similar and equivalent to quiet. Therefore, there is a possibility that for human speech and vocoded speech, in the -1dB condition, listening effort is infact not being indexed as we had initially thought.

Furthermore, when comparing the -1dB condition and -3dB condition, we observe that Fastspeech 2 and Merlin also behave differently to all other systems. For example, in the linear term, Fastspeech 2 is the only system that increased in slope steepness and in the cubic terms, those systems were the only ones that increased in slope steepness. These observations align with what is expected when cognitive load is measured and therefore it seems plausable that for the poor

performing systems in this evaluation (Fastspeech 2 and Merlin), a sufficient load was exerted in -1dB SNR condion. Therefore, listening difficulty for only these 2 systems were being indexed in the -1dB condition. However, for the other high quality systems, the pupil response is perhaps indexing levels of engagement as observed in the quiet condition in previous chapters. Although, Merlin has a greater response for -5dB than -3dB, it is still likely that ceiling capacity is reached at the -3dB condition as it is unlikely that Merlin would demand the same cognitive load for -3dB and -5dB as shown in Figure 7.5. In Chapter 6, Merlin was shown to reach ceiling capacity at -3dB and therefore since the architecture of the system between Chapter 6 and 7 remained the same, it is more likely that ceiling at -3dB was reached in this experiment too. Therefore, levels of engagement are being indexed for human, Multi-band melGAN and Tacotron 2. This finding also explains why the ERPD for vocoded speech in -3dB is much lower than the ERPD in -1dB.

In light of this, in the -1dB condition, vocoded speech demands a higher level of engagement to human speech and Tacotron 2. This is consistent across all terms. It has the highest mean, steepest rising and falling slope and sharpest peak in comparison to human speech and Tacotron 2. Human speech and Tacotron are found to be equivalent in most terms and therefore demand a similar level of engagement. Comparing Fastspeech 2 and Merlin, it is evident that Fastspeech 2 is more difficult to process than Merlin in the -1dB SNR. This is surprising as Fastspeech 2 uses Transformers as is said in the literature to have comparative quality with current state-of-the-art models.

Finally, comparing the ERPD responses for Human, Multi-band melGAN and Tacotron 2 in the -3dB and -5dB condition, we observe that there is an increase between the conditions for human and vocoded speech whilst there was a very small difference for Tacotron 2. Tacotron 2 evokes a higher ERPD response in the -3dB condition than vocoded and human speech which explains why the gap is small. This finding together with the analysis of the time terms suggests that Tacotron 2 still seems to be more difficult to process than human and vocoded speech specifically in the -3dB condition and as a result demands more cognitive load than human and vocoded speech.

Summary

The cognitive load of synthetic speech produced by the latest state-of-the-art high quality text-to-speech systems were investigated in this chapter. This was performed by measuring the evoked pupil response when listening to speech produced by these systems in the presence of noise, specifically -1dB, -3dB and -5db SNRs.

All systems scored above the expected WER thresholds for the -1dB (20%) except for Merlin. This is particularly interesting because in the -1dB condition in Chapter 6, all systems had WERs below 15% including Merlin and thus remained close to ceiling. The difference between Chapter 6 and 7 is (1) a different speaker was used, (2) American English instead of British English was used and (3) different sentence material was used (ie. longer sentences). Therefore, it is clear that all these factors play a role in the WER results achieved.

For the self-reported measures, we observed that human speech was more natural than vocoded

only in the -1dB SNR but equivalent in the harder SNRs. Merlin was found to be the least natural. However, as the SNR decreases we see that the naturalness perceptions begin to converge and listeners find it more difficult to separate the true naturalness differences between the systems. In terms of self-reported cognitive load, a similar trend is observed. Therefore, evaluating TTS in the presence of noise interferes with subjective measures collected by listeners and thus makes them somewhat unreliable. The pupil response results unraveled new findings that were initially not considered before. We found that the pupil response is directly influenced by the amount of load exerted on the listener. In conditions where the overall load is insufficient we observe that the pupil appears to be indexing something other than listening difficulty as previously observed in quiet. This is due to coefficients in the various time terms being inconsistent with what we expect. Therefore, when conducting such experiments it is critical to ensure that comparisons are made amongst varying SNR conditions in order to determine if the trend of results are consistent we what is expected. The pupil response results revealed that Merlin and Fastspeech 2 are the poor performing TTS systems in this experiment as both these systems appear to reach ceiling in the -3dB condition. It was unexpected to observe that Fastspeech 2, which is considered to be a state-of-the-art model performed so poorly in relation to Tacotron 2. We observed that vocoded speech and human speech behave similarly which leads us to believe that vocoded speech and human speech demand the same amount of cognitive load. Since the vocoder uses features extracted directly from human speech whilst the TTS model predicts these features this therefore means that the contributions to an increased cognitive load in TTS systems are a result of the predicted features generated by the acoustic model. Although Tacotron 2 was not far behind vocoded and human speech, it was shown that it still demands more cognitive load than human and vocoded speech. Given that we have identified that Fastspeech 2 versus Tacotron 2 perform so differently, this information is valuable as it allows us to investigate further where the contributions in the model specifically stem from. For example, by understanding the differences in the acoustic model between Tacotron 2 and Fastspeech 2 will enable us to dive a bit deeper in understanding where exactly the increased load comes from in Tacotron 2 so that we can develop better TTS models that demand the same amount of cognitive load as human speech, or if not, even better in future. In conclusion, similarly to what was observed in Experiment 1, speech produced by state-of-the-art TTS is moving in the direction of demanding a similar amount of cognitive load as human speech. In Experiment 1, we observed that DC-TTS in relation to Tacotron 2 demands more cognitive load and in this experiment we observed that Fastspeech 2 in relation to Tacotron 2 also demands more cognitive load. Therefore, investigating the differences between these 3 systems will enable us to further understand the source of increased cognitive in current state-of-the-art text-to-speech systems.

Chapter 8

Summary of investigations in Part II

In Part II of this thesis, the aim was to apply pupillometry as a method for measuring the cognitive load of synthetic speech to current state-of-the-art TTS models as the systems that were initially tested in Part I are no longer state-of-the art. Text-to-speech systems improved drastically since the introduction of deep neural networks (DNNs). Therefore in Part II, the aim was to determine whether the hypothesis that speech produced by TTS is more difficult to listen to than human speech still holds true for modern TTS systems that are trained using deep neural networks. In addition, we take this a step further in Chapter 6 by using the pupillometry paradigm as a method to discover where the contributions to increased cognitive load come from. The key findings of this investigation lead us to the final experiments in Chapter 7 that investigated sequence-to-sequence based TTS models. Since there was a 3 year break in this thesis between Chapter 7 Experiment 1 and Experiment 2, it was important to test the methods derived in this thesis on the latest state-of-the-art systems to ensure that the work produced in this thesis remains relevant.

8.1 Discussion and Concluding Remarks

Cognitive load of DNN-based speech synthesis

The pupillometry paradigm in Part I was used to investigate the cognitive load of DNN-based speech synthesis. In doing so, we set up the experiments in such a way that could aid us in understanding where the contributions to increased cognitive load of a DNN-based speech synthesis system comes from. This was performed by investigating the cognitive load of each vocoder speech parameter modelled in the chosen DNN system. In these experiments, the systems evaluated were configurations that stepped from human speech to a full DNN TTS system. The in-between configurations were entirely artificially created by swapping out predicted features with features extracted directly from human speech.

In Part I, we identified that the properties associated with low listening effort include, low mean, a broad peak and a rising and falling slope with a low gradient whilst the opposite properties are associated with high listening effort. We also established that when an insufficient amount of load is placed on the listener, such as listening in quiet, then attention resources are more likely to be

indexed and therefore what we observe is levels of engagement rather than listening difficulty. As a reminder, attention resources indexed as a result of high levels of engagement can be seen as both positive and negative. Positive in that listeners choose to allocate more attention to listen to speech as a form of engaged interest and motivation, and negative in that listeners are forced to pay closer attention to speech that is challenging to process.

In quiet, the WER results alone offered no insightful information as all WERs across all configurations were not significantly different to one another. With regards to self-reported cognitive load, human speech was the easiest to listen to whilst TTS was the hardest. The pupil responses, however, showed contrasting findings to those reported in the self-reported measures where human speech evoked a pupil response with properties that are associated mostly with high listening effort (high mean, steep rising and falling slopes). Since we know that it is unlikely for TTS to be easier to listen to than human speech, we concluded that in quiet, the pupil response when listening to human speech is indexing more the level of engagement rather than listening difficulty. We also observed that the system where correlations between the acoustic features remain intact was found to have increased levels of engagement compared to the systems in which the correlations were destroyed. Therefore, features predicted separately lead to a reduced level of engagement. In this experiment, we also observed that when listening to speech produced by a DNN-speech synthesizer human and synthetic speech reflect similar properties whereas the older TTS systems in Part I had shown contrasting properties, which confirms that even though levels of engagement aren't as high as human speech, synthetic speech is getting closer to human speech.

When listening in noise, intelligibility was still quite high even at the most difficult SNR level. This shows that DNN-based speech synthesis still performs reasonably well even under adverse conditions. For the self-reported scores, we observe that as the SNR decreases, listeners tend to lose their ability in differentiating between the different systems and therefore their scores become less reliable. This therefore motivates the need for using physiological measures such as pupillometry that are more reliable.

For the pupil responses, we observed that human speech and vocoded speech are processed similarly in the -1dB condition. We also later discovered that the values in the intercept in the -1dB condition are all less than the values in the quiet condition except for the TTS system. Therefore, it is plausible that listening difficulty is only being measured for some of the conditions in the -1dB condition. We also observe that the TTS system evokes a smaller pupil response, has a less steep slope in both the linear and cubic terms in the -5dB condition, and as previously explained, this is likely due to the TTS system reaching ceiling capacity in the -3dB condition.

The key findings of this experiment showed that DNN-based speech synthesis, specifically the Merlin system evaluated, was more difficult to listen to than human speech. Vocoded speech on the other hand, was found to be processed in a similar manner to human speech. Therefore, the cognitive load contributions appear to stem from the acoustic and duration models. By comparing the various configurations, we further discovered that poor mel-cepstral coefficient prediction and duration prediction were the leading factors that contributed to an increased cognitive load. In addition, by combining predicted features with natural features, dependencies between these features were

deliberately destroyed and in doing so, we discovered that having independently predicted features also resulted in an increased cognitive load. This finding was important because in traditional DNN-based speech synthesis systems, such as Merlin, features are predicted separately and then passed individually to a vocoder to generate the output waveform. Producing speech in this way is more cognitively demanding than human speech.

Fortunately, modern TTS systems such as sequence-to-sequence based TTS have done away with the approach of modelling features separately. Nowadays, end-to-end solutions are common and therefore most features are predicted simultaneously in a single unified framework. Therefore, it became important to evaluate sequence-to-sequence based TTS in this thesis to understand whether having a unified framework does infact reduce the cognitive load of text-to-speech.

Setting up the experiment by teasing apart the different components certainly made a difference in helping us understand the flaws in the development of text-to-speech systems. Flaws that could not have been discovered from using intelligibility and naturalness tests alone. Therefore, pupillometry is a useful tool to consider when wanting to delve deeper into understanding the shortcomings of TTS in relation to human speech, not just from the perspective of cognitive load but also to gauge the overall performance of the system itself.

The cognitive load of state-of-the-art TTS

In these experiments, we were particularly interested in understanding whether cognitive load of state-of-the-art TTS still demand higher cognitive load than human speech. Therefore we evaluated two of the most recent state-of-the-art TTS systems namely Fastspeech 2 and Tacotron 2. In addition, we included the Merlin system from Chapter 6 for comparison. By including the vocoder alone, this too also aided us in understanding where cognitive load contributions - if any - come from.

With regards to intelligibility, Merlin has the highest WER which was expected as it was specifically included for comparison and we expected it to perform the worst. Surprisngly, Fastspeech 2 was found to be just as bad as Merlin, especially in the more challenging SNRs. With regards to self-reported naturalness, although human speech generally was rated the most natural, in the more challenging SNRs these converged more with the other systems. Only Merlin and Fastspeech 2 were consistently rated the least natural. As we had observed several times in this thesis, naturalness ratings become less reliable under noisy conditions. Therefore, if one expects to evaluate TTS in noise which is more reflective of real-world conditions, then naturalness scores will become useless. With regards, to the self-reported cognitive load, similar results were obtained. Merlin and Fastspeech 2 were perceived to be most difficult to listen to across all experiments, whilst the other systems and human speech were easier but equivalent. Therefore, the self-reported cognitive load scores also become less reliable, which emphasises the need to look beyond traditional evaluation methods.

With regards to the evoked pupil responses, we could immediately identify a difference in processing between the high quality and low quality speech synthesizers. Merlin and Fastspeech 2 in the easier SNR demanded the most cognitive load. However, there was reason to believe that for

all other systems, an insufficient load was placed on the listener. Parameter estimates for all other systems were not aligning with what we would have expected to see, the parameter estimates were lower in the easier SNR compared with the more difficult SNR. When comparing the remaining systems in the more challenging SNRs, we observed that human and vocoded speech behave similarly. This finding was also observed in Chapter 6. Therefore, all vocoders investigated in this thesis were good enough to demand a similar cognitive load to human speech. Even though neural vocoders are much better than the WORLD vocoder for instance, the fact that we observe the same trend in Chapter 6 indicates that the vocoder is not the problem in the text-to-speech synthesis systems. Interestingly, although Tacotron 2 demanded more cognitive load than human and vocoded speech in the more challenging SNRs, in both its versions it was not too far off. From all three state-of-the art TTS systems ie., Tacotron 2, DC-TTS and Fastspeech 2, Tacotron 2 outperformed DC-TTS and Fastspeech 2 in terms of demanding the least amount of cognitive load. Although it was out of the scope of this thesis to investigate this further, these findings could at the very least point us to the aspects of the architecture that is contributing to an increased cognitive load. For instance, the key difference between Tacotron 2 and Fastspeech 2 is the model itself. Perhaps transformers are not a good as we think they are compared to more auto-regressive approaches such as Tacotron 2. Comparing Tacotron 2 with DC-TTS which are both auto-regressive, a key difference is that DC-TTS works with linear spectrograms whilst Tacotron 2 works with mel-spectrograms and therefore perhaps our cognitive processing systems seem to have a harder time processing linear spectrograms than mel-spectrograms. Another plausible contribution could be in the way in which alignments are implemented or the fact that Tacotron 2 uses RNNs whilst DC-TTS uses CNNs. Overall, the findings of this chapter prove that the cognitive load for synthetic speech is still greater than human speech despite the self-report measures telling us otherwise. Therefore measuring the cognitive load of synthetic speech is still relevant and pupillometry has been proven to be useful in detecting differences even between high quality speech synthesizers. Furthermore, such a technique, if set-up correctly, like we did in Chapter 6, can be used to delve deeper into uncovering insights that traditional methods may not offer.

Chapter 9

Conclusions and future work

9.1 Contributions

The primary contribution of the work presented in this thesis is a methodology for measuring the cognitive load of synthetic speech. Traditional tests that are currently used to evaluate text-to-speech typically comprise of evaluating only the naturalness and intelligibility. However, such methods can be considered limiting as it does not give us a deeper understanding of how synthetic speech is processed by the human cognitive processing system. In the literature it has been shown that synthetic speech is considered to be more cognitively demanding when listened to compared to human speech. This is an important factor to consider during text-to-speech evaluations, as synthetic speech is becoming increasingly used in real-world applications and therefore the user experience also becomes equally important. Understanding the cognitive load of synthetic speech is important for two key reasons: (1) if synthetic speech is more cognitively demanding than human speech this can lead to negative implications such as fatigue especially if our cognitive processing system is placed under high load for extended periods of time. For example, in the use case of listening to an audio-book that uses text-to-speech, (2) by understanding the differences between human and synthetic speech could aid us in understanding where increased contributions come from and therefore can equip us with the knowledge we require to optimise synthetic speech for low cognitive load.

Research Question 1a: How can we measure the cognitive load when listening to synthetic speech?

In Part I, we investigated two methodologies for measuring the cognitive load of text-to-speech systems. First, the dual-task paradigm was investigated. The dual-task paradigm was used to measure the cognitive load of synthetic speech in the past. However, the key difference between the work already done and ours is that the cognitive load of only rule-based speech synthesis systems were previously evaluated. Text-to-speech synthesis systems has drastically improved since then and therefore it became important to understand how more recent text-to-speech systems interact with our cognitive processing system. The second method investigated was pupillometry. Pupillometry is

a method that already exists in the literature. However, to our knowledge, ours is the first attempt to measure the cognitive load of synthetic speech using pupillometry. Therefore, the work presented is this thesis was a starting point for determining whether pupillometry is a viable measure for this purpose.

Research Question 1b: Can a suitable paradigm be developed that is capable of detecting differences in cognitive effort required to listen to various TTS systems?

The dual-task paradigm was first investigated as a potential method for measuring the cognitive load of synthetic speech. However, it was hard to control what participants do during the experiment. Cognitive load, in the dual task paradigm, is measured indirectly through the deterioration in performance in the secondary task. But knowing whether the deterioration is purely due to cognitive load was difficult to determine. There was no way to be certain that participants were prioritising the primary task, or whether a training effect played a role in them becoming faster in the secondary task and lastly, there was uncertainty whether sufficient load was placed on the listener. It seemed likely, that the participants could manage to perform both tasks at high performance without them deteriorating in performance in the secondary task. Therefore, we concluded that the dual-task paradigm was unreliable and not suitable for our purposes.

Pupillometry was then investigated as the pupil response has been shown to index listening effort in many studies. In the first experiment, we replicated the traditional intelligibility test by utilising SUS in the listening task for pupillometry. The pupil response is sensitive to semantic information and it was previously found that it is more effortful to process difficult and/or more complex sentences such as SUS than easier semantically correct sentences such as SMS [Beatty, 1982]. In our experiment, differences in the speech synthesizers compared were detected. However, we observed that for natural speech the peak latency was delayed when listening to SUS compared to listening to SMS. This lead us to believe that SUS and SMS are not processed the same by our cognitive processing system. Therefore, if we want to understand the true reflection of the cognitive load imposed that matches the real-life scenario of listening to text-to-speech, it is more ecologically valid to use SMS rather than SUS as the cognitive processing system may not be using the same mental resources.

We then investigated the influence of the pupil response when listening in quiet to SMS. We were concerned that an insufficient load would be placed on the listener and therefore we would obtain inaccurate results. This concern turned out to be correct. Results showed the opposite to what was expected where the highest quality speech synthesizer was shown to demand the most listening effort which was unlikely. Therefore, it became clear that when listening in quiet, the human cognitive processing system was indexing some other resource and not the working memory which was the resource we wished to index in this thesis. This finding was validated as the same finding was observed across several experiments and it was consistently shown that when listening in quiet

the pupil response was greater than when listening in noise which is impossible thereby confirming that listening difficulty was not being indexed. In addition, human speech behaved differently to synthetic speech and therefore there was concern raised around whether human speech and synthetic speech are comparable in this experiment. Since one of the important reasons for evaluating cognitive load is to understand the difference between human and synthetic speech, if they are not comparable within a given experiment then the results of that experiment become meaningless to us.

To ensure sufficient load was placed on the listener, we investigated the pupil response when listening in the presence of noise. Results aligned with what we expected and therefore it was confirmed that the resource we were indexing was the working memory which was the resource we wished to measure. Listening difficulty differences between human and synthetic speech were detected. Apart from understanding the cognitive differences between human and synthetic speech, one other important requirement we were looking for when selecting the most suitable methodology was that the chosen method needed to also detect listening difficulty differences between the various TTS systems evaluated. Results in our experiments showed that the pupil response is sensitive to changes in speech quality and therefore differences between the speech synthesizers were also detected.

Research investigating listening effort identified the dominant cognitive processes for speech processing as working memory, speed of processing, linguistic knowledge and attention [Gagne et al., 2017]. Throughout this thesis, these processes became crucial in understanding what the pupil was indexing. It was out of the scope of this thesis to understand exactly which process was indexed, but it became more clearer that the set-up of the pupillometry experiment is extremely important for ensuring that we are only indexing the cognitive process we desire.

High listening effort is associated with high means, steep rising and falling slopes and sharp peaks. From the results we observed, on one hand, high listening effort in quiet was associated more with high levels of engagement which can be seen as both positive and negative. Positive in that high listening effort could mean that the speech is more engaging and therefore more attention resources are demanded and negative in that more attention is needed to focus when the speech is more challenging to listen to. On the other hand, high listening effort in noise is associated more with listening difficulty. In other words, properties of high listening effort indicate that a system is more difficult to listen to than systems that have properties associated with low listening effort. Since listening difficulty in noise is naturally difficult, we instead made comparisons in terms of which systems were more cognitively demanding than another rather than differentiating between high and low listening difficulty.

Therefore the most suitable method to detect difference in cognitive effort between TTS systems is to apply the pupillometry paradigm using semantically meaningful sentences that are masked with noise.

Research Question 2: Does synthetic speech demand greater cognitive effort compared to human speech?

In all cases, we found that human speech demanded the least amount of cognitive load even when listening in the presence of noise. Vocoded speech was in most cases found to be equivalent to human speech. High quality synthesizers demanded more cognitive resources than human and vocoded speech even when evaluating the most recent state-of-the-art TTS systems. In earlier chapters, we started with older TTS systems and gradually moved towards more recent TTS systems. It was evident, that earlier systems demanded more cognitive load than more recent TTS systems. More importantly, we observed that state-of-the-art TTS systems such as Tacotron 2 are actually not far behind from human and vocoded speech with respect to the amount of cognitive load demanded. Interestingly, two other state-of-the-art systems which have been shown to produce high quality TTS output demanded significantly more cognitive load than Tacotron 2. Infact, they were shown to be too difficult to manage in the more challenging SNRs. This finding was useful, as it tells us that even though these systems perform really well when listened to in quiet, they fall short in the presence of noise and therefore if embedded into real-world applications this may result in an unpleasant experience for the user.

Research Question 3: If synthetic speech demands greater cognitive effort compared to human speech, what are the contributing factors that lead to an increased cognitive load in synthetic speech processing?

By going one step further, we used pupillometry as a way to delve deeper into understanding where the contributions comes from. Although, self-reported measures help us gauge the overall performance of the system they do not help us understand why the system performs the way that it does. By creating stimuli that stepped from human speech to a full DNN TTS system, we were able better understand what aspects of the system lead to an increased cognitive load. Within the Merlin DNN speech synthesis system, we discovered that an increased cognitive load is due to poor spectral feature prediction and poor duration modeling. We also discovered that by predicting spectral and excitation features separately breaks the correlation that exists between them and this too resulted in an increased cognitive load. This motivates extending evaluation methods to new measurements like pupillometry that have the potential to provide much more meaningful insights than a 1-dimensional score obtained in a self-reported assessment.

Research Question 4: With the recent advancements made in TTS, do modern TTS systems still demand greater cognitive load than human speech?

The latest TTS systems have already overcome these increased contributions. Recent TTS systems typically have a single unified framework and therefore all features are predicted simulta-

neously. However, even with these improvements we still observed that synthetic speech is more challenging to listen to that human speech. However, it is clear that the demand between human speech and synthetic speech is reducing.

Research Question 5: Is the research conducted in this thesis still relevant?

Pupillometry has been proven to be a viable method for measuring the cognitive load of synthetic speech. It is capable of detecting differences between human and synthetic speech, detecting differences between low quality and high quality TTS systems and also detects differences between current state-of-the-art TTS systems. Therefore, this method still remains relevant even for the most recent state-of-the-art TTS models. A disadvantage of such a method however is that it requires an in-depth understanding of analysing pupil data in order to extract meaningful information. It is also a long and tedious method compared to traditional methods that are much quicker. Nevertheless, it is extremely important for us to understand the cognitive load of TTS for the purpose of ensuring that there is no negative implications placed on the end-user when listening to TTS. Therefore, whilst it may not be the fastest or cheapest method to measure the cognitive load of synthetic speech, it has been shown to work effectively and provide meaningful results that can aid us in developing better TTS systems.

9.2 Lessons learnt

- Most current evaluation methods are conducted in ideal (quiet) listening conditions, yet end-users in a realistic scenario won't always be listening to TTS under such conditions. Therefore there is a need to move towards more realistic evaluation strategies as motivated by the results of the two state-of-the-art TTS systems that were shown to demand a significantly greater cognitive load than other TTS systems. This provides further motivation for evaluating TTS in the presence of noise. Thereby making our methodology a suitable and practical method.

- According to [Winn et al., 2018], the pupil response is non-linear and a small pupil dilation can be evoked not only when performing an easy task but also for a hard task when effort is voluntarily withdrawn by the listener. For example, in the case of the listener experiencing fatigue. Therefore the pupil response is closely related to a listeners willingness to listen. In our experiments, this finding was confirmed. In several experiments, we observed a smaller evoked pupil response than expected when listeners were listening to speech produced by poorer quality TTS systems. Thus, we interpreted this finding as the listener reaching ceiling capacity as a consequence of withdrawing because the task was too difficult to manage. This helped us differentiate the high quality synthesizers from the low quality synthesizers. As mentioned, it was surprising to see that even state-of-the-art TTS systems reached ceiling in harder SNRs whilst other TTS systems were still manageable.

- When listening in the presence of noise, the self-reported measures were shown to become less

discriminative as scores between the various speech synthesizer and human speech converged as the SNR decreased. However, the pupil responses evoked was still able to shed light on differences in the various systems in these SNR conditions. Therefore, there is benefit is performing evaluations with pupillometry compared to relying only on self-reported measures especially when evaluating the cognitive load in the presence of noise.

9.3 Future Work

Many interesting findings surfaced when conducting the experiments in this thesis, but it was out of the scope of this thesis to further explore. This, however creates space for future work as follows:

- Although the dual task paradigm did not prove useful, one could investigate repeating the experiments conducted in this thesis but instead setting the primary task as a listening task in the presence of noise which should solve the insufficient load problem as it did when using pupillometry.
- In quiet, we observed that a different cognitive process was being indexed. Future work can investigate this further to understand exactly which cognitive resource is being indexed in these conditions.
- In this thesis, we used speech shape-noise as a starting point. However, since the start of this work, research effort has been placed on producing more realistic noise [Chermaz et al., 2019]. Therefore, it will be interesting to evaluate the cognitive load of text-to-speech in the presence of more realistic noise that is a truer reflection of real-world conditions. For example, noise in a cafeteria or train station.
- In the last chapter, we discovered differences in cognitive load between state-of-the-art TTS systems. Therefore, taking this a step further like we did in Chapter 6 for the DNN-based-speech synthesizer, would be meaningful to better understand exactly where increased cognitive load contributions come from in sequence-to-sequence based TTS systems and transformer-based TTS systems.

Appendix A

Test Sentences used in Chapter 6

Semantically unpredictable sentences taken from 2011 Blizzard Challenge used as test set for Chapter4, Experiment 1A:

- 0001 The hand finished over the cold leg.
- 0002 The book dragged the seem that slept.
- 0003 The sharp course paid the heart.
- 0004 Why does the jar build the white area?
- 0005 Catch the camp and the trade.
- 0016 The machine rushed since the red dime.
- 0020 Want the form and the kiss.
- 0010 Obey the force or the wiff.
- 0008 The mad soup promised the relief.
- 0015 Command the farm or the stress.
- 0009 How does the spoon tie the capital mind?
- 0018 The deep zoo harmed the house.
- 0019 Where does the pain seek the happy sense?
- 0013 The fat wind found the list.
- 0012 The drink gained the sheet that coped.
- 0011 The face posed after the dry plane.
- 0017 The friend trimmed the horse that escaped.
- 0007 The text cleared the library that worked.

- 0006 The view ducked as the fast bank.
- 0014 When does the dance question the loud clock?
- 0021 The fruit cried against the strong guitar.
- 0022 The end bought the earth that came.
- 0023 The polite top selected the design.
- 0024 How does the breakfast cast the new heat?
- 0025 Hear the meal or the smile.
- 0030 Lift the sight and the salt.
- 0040 Scare the laugh or the money.
- 0039 When does the bird fold the wrong plant?
- 0029 Where does the cup trust the tall field?
- 0026 The cat gazed in the light air.
- 0032 The day sliced the fear that wept.
- 0027 The joke raised the art that talked.
- 0028 The brief dust touched the test.
- 0035 Owe the date and the square.
- 0031 The sun crawled to the green music.
- 0037 The bridge felt the space that grew.
- 0034 Why does the store greet the strange road?
- 0036 The car died above the pure war.
- 0033 The young note hired the length.
- 0038 The clean child saved the growth.
- 0041 The role slipped behind the wild nose.
- 0042 The star cut the wall that prayed.
- 0043 The full choice mixed the tree.
- 0044 How does the snow drain the terrible bag?

- 0045 Help the age or the search.
- 0047 The rest kicked the rain that crept.
- 0055 Notice the light and the religion.
- 0053 The whole tone punished the word.
- 0056 The job danced up the fresh voice.
- 0059 Where does the help hold the pale range?
- 0054 When does the land judge the bad potato?
- 0049 Why does the cucumber get the thick song?
- 0051 The cell coughed out the wet dirt.
- 0048 The broad reep heared the case.
- 0046 The amount tried with the wide lunch.
- 0050 Paint the bar and the meat.
- 0058 The huge man pushed the key.
- 0052 The boy wiped the apartment that stared.
- 0060 Keep the head or the telephone.
- 0057 The oil fixed the group that argued.
- 0061 The staff ached from the gray lot.
- 0062 The milk wrote the time that sang.
- 0063 The true desk marked the mess.
- 0064 Where does the way open the vast knee?
- 0065 Guide the health and the doubt.
- 0072 The wood poured the tooth that fell.
- 0077 The home described the month that marched.
- 0070 Pull the truth or the boot.
- 0078 The rich tax caused the brain.
- 0076 The plan killed on the good boat.

- 0074 How does the gas cut the high lock?
- 0079 When does the boss toss the nice fire?
- 0069 Why does the claim report the sad string?
- 0066 The speech stayed through the clear hall.
- 0075 Need the sort and the coat.
- 0071 The lake leapt of the honest step.
- 0067 The flow launched the goal that dreamed.
- 0068 The free head missed the race.
- 0073 The old shape attacked the shoe.
- 0080 Join the card or the clay.
- 0081 The glass rose by the sick brush.
- 0082 The toe liked the escape that cared.
- 0083 The blue mistake soaked the lip.
- 0084 Where does the film smell the alert side?
- 0085 Heed the truck or the neck.
- 0088 The regular bread feared the kind.
- 0092 The press joined the water that clung.
- 0094 Why does the gift unlock the large floor?
- 0086 The person walked down the brown cost.
- 0090 Answer the town and the mouth.
- 0095 Bring the piece or the plug.
- 0087 The strength grasped the love that won.
- 0100 Hit the part and the page.
- 0096 The sound stood near the dark breath.
- 0098 The perfect school dared the band.
- 0091 The loss listened before the low chair.

- 0089 How does the eye use the important front?
- 0093 The warm sugar loved the body.
- 0099 When does the style make the slow weather?
- 0097 The party brushed the firm that appeared.

Semantically unpredictable sentences taken from 2011 Blizzard Challenge used as test set for Chapter4, Experiment 1B:

- 0001 The word leaned through the strange mouth.
- 0002 The glass poured the date that cared.
- 0003 The thin sign launched the sort.
- 0018 The fair line saved the hall.
- 0007 The hand caused the export that fell.
- 0012 The space sold the ball that laughed.
- 0004 Why does the ground rent the daring style?
- 0009 How does the the hot job?
- 0020 the gun or the door.
- 0011 The you fell under the dead day.
- 0019 Where does the mind feel the straight tree?
- 0014 When does the trade toss the quick dust?
- 0016 The page slept for the big chance.
- 0005 Warn the room or the piece.
- 0015 Use the form and the sound.
- 0006 The fact prayed by the pure ship.
- 0008 The blue search brushed the suspect.
- 0017 The string dropped the chair that wept.
- 0010 Keep the clay and the court.
- 0013 The clear price kept the stage.
- 0021 The song duck the happy name.

- 0022 The bank pleased the tax that spoke.
- 0023 The good lot dared the frame.
- 0024 Where does the view note the tall trip?
- 0034 How does the kid make the real neck?
- 0038 The great home wiped the part.
- 0032 The work marked the shot that coughed.
- 0029 When does the end add the free block?
- 0028 The whole team fixed the flow.
- 0037 The box traced the fear that smiled.
- 0036 The tone came to the brown board.
- 0035 Guide the leg or the score.
- 0025 Take the fire or the child.
- 0026 The cried in the super cell.
- 0033 The old week meant the force.
- 0030 Get the sort and the rest.
- 0031 The house crawled from the fine bill.
- 0039 Why does the wine bless the deep car?
- 0027 The theme dealt the month that coped.
- 0040 Spend the club and the year.
- 0041 The faith dangled on the wide range.
- 0042 The text heard the horse that rose.
- 0043 The gray kind grabbed the heart.
- 0050 Trust the pool or the friend.
- 0054 When does the top thank the open truck?
- 0057 The store gained the case that died.
- 0044 When does the jazz find the sweet dance?

- 0052 The foot cured the class that posed.
- 0046 The roof dwelt down the firm.
- 0047 The air built the war that ceased.
- 0059 How does the blood hang the bad plant?
- 0045 Solve the desk and the earth.
- 0053 The sure groups the size.
- 0056 The law stood with the pale staff.
- 0055 Bite the rate and the girl.
- 0058 The hard test shared the field.
- 0060 Convict the hair or the wife.
- 0048 The green book killed the doubt.
- 0049 Why does the head send the cold wind?
- 0051 The change stayed with the wild will.
- 0061 The world marched through the nice goal.
- 0062 The task grasped the sky that leapt.
- 0063 The short phase dragged the farm.
- 0066 The floor clung in the press.
- 0069 Why does the man lift the small shape?
- 0070 Cast the front and the mass.
- 0065 Treat the length or the choice.
- 0067 The state paid the charge that talked.
- 0078 The brief street blamed the gas.
- 0071 The sense went the dark boat.
- 0073 The wrong bit liked the film.
- 0072 The stress owned the bar that crept.
- 0064 When does the arm need the long claim?

- 0075 Seek the loss or the strength.
- 0068 The sharp permit mixed the face.
- 0079 How does the time waste the black type?
- 0076 The knife looked for the boring abstract.
- 0074 Where does the light want the true drink?
- 0080 Throw the bridge and the cost.
- 0077 The plane held the town that gazed.
- 0081 The speech walked by the rich death.
- 0082 The road cleaned the way that
- 0083 The large growth trimmed the oil.
- 0094 Why does the boy abstract the poor game?
- 0087 The church earned the truth that lived.
- 0085 Tie the bed or the
- 0093 The young showed the voice.
- 0088 The wide stock joined the post.
- 0097 The hill the rain that lied.
- 0084 When does the love tale the fresh night?
- 0090 Bear the land and the rule.
- 0100 Export the and the planned.
- 0089 How does the art hit the brutal point?
- 0095 Suspect the side or the race.
- 0091 The role stared to the vast life.
- 0096 The son paused from the sick course.
- 0086 The age roamed down the bride list.
- 0098 A soft place raised the speed.
- 0092 The pain spared the thing that screamed.

- 0099 Where does the camp greet the new square?

Semantically meaningful sentences taken from 2011 Blizzard Challenge used as test set for Chapter4, Experiment 2 and Experiment 3:

- 419 The point of the steel pen was bent and twisted.
- 420 There is a lag between thought and act.
- 421 Seed is needed to plant the spring corn.
- 434 These coins will be needed to pay his debt.
- 430 Screen the porch with woven straw mats.
- 431 This horse will nose his way to the finish.
- 424 The chap slipped into the crowd and was lost.
- 433 He picked up the dice for a second roll.
- 438 The smell of burned rags itches my nose.
- 432 The dry wax protects the deep scratch.
- 422 Draw the chart with heavy black lines.
- 428 Say it slowly but make it ring clear.
- 425 Hats are worn to tea and not to dinner.
- 426 The ramp led up to the wide highway.
- 437 The vamp of the shoe had a gold buckle.
- 423 The boy owed his pal thirty cents.
- 429 The straw nest housed five robins.
- 436 Twist the valve and release hot steam.
- 435 The nag pulled the frail cart along.
- 427 Beat the dust from the rug onto the lawn.
- 439 New pants lack cuffs and pockets.
- 440 The marsh will freeze when cold enough.
- 441 They slice the sausage thin with a knife.
- 443 A grey mare walked before the colt.

- 449 The desk and both chairs were painted tan.
- 451 A clean neck means a neat collar.
- 450 Throw out the used paper cup and plate.
- 452 The couch cover and hall drapes were blue.
- 457 The cleat sank deeply into the soft turf.
- 458 The bills were mailed promptly on the tenth of the month.
- 444 Breakfast buns are fine with a hot drink.
- 445 Bottles hold four kinds of rum.
- 456 Turn out the lantern which gives us light.
- 454 The wall phone rang loud and often.
- 448 Drop the ashes on the worn old rug.
- 446 The man wore a feather in his felt hat.
- 453 The stems of the tall glasses cracked and broke.
- 447 He wheeled the bike past the winding road.
- 455 The clothes dried on a thin wooden rack.
- 442 The bloom of the rose lasts a few days.
- 459 To have is better than to wait and hope.
- 460 The price is fair for a good antique clock.
- 461 The music played on while they talked.
- 463 The bunch of grapes was pressed into wine.
- 473 The tin box held priceless stones.
- 475 The case was puzzling to the old and wise.
- 471 The kite flew wildly in the high wind.
- 474 We need an end of all such matter.
- 472 A fur muff is stylish once more.
- 478 The youth drove with zest, but little skill.

- 462 Dispense with a vest on a day like this.
- 465 The hinge on the door creaked with old age.
- 464 He sent the figs, but kept the ripe cherries.
- 468 Thick glasses helped him read the print.
- 466 The screen before the fire kept in the sparks.
- 467 Fly by night and you waste little time.
- 476 The bright lanterns were gay on the dark lawn.
- 470 The chair looked strong but had no bottom.
- 469 Birth and death marks the limits of life.
- 477 We don't get much money but we have fun.
- 479 Five years he lived with a shaggy dog.
- 480 A fence cuts through the corner lot.
- 481 The way to save money is not to spend much.
- 491 A man in a blue sweater sat at the desk.
- 497 A force equal to that would move the earth.
- 492 Oats are a food eaten by horse and man.
- 486 Send the stuff in a thick paper bag.
- 496 Tuck the sheet under the edge of the mat.
- 488 They told wild tales to frighten him.
- 489 The three story house was built of stone.
- 494 A sip of tea revives his tired friend.
- 487 A quart of milk is water for the most part.
- 498 We like to see clear weather.
- 482 Shut the hatch before the waves push it in.
- 484 Crack the walnut with your sharp side teeth.
- 490 In the rear of the ground floor was a large passage.

- 493 Their eyelids droop for want of sleep.
- 485 He offered proof in the form of a large chart.
- 483 The odour of spring makes young hearts jump.
- 495 There are many ways to do these thing.
- 499 The work of the tailor is seen on each side.
- 500 Take a chance and win a china doll.
- 501 Shake the dust from your shoes, stranger.
- 508 The water in this well is a source of good health.
- 502 She was kind to sick old people.
- 503 The square wooden crate was packed to be shipped.
- 510 That guy is the writer of a few banned books.
- 509 Take shelter in this tent, but keep still.
- 516 The pleasant hours fly by much too soon.
- 506 Smile when you say nasty words.
- 517 The room was crowded with a wild mob.
- 505 We dress to suit the weather of most days.
- 512 The door was barred, locked, and bolted as well.
- 511 The little tales they tell are false.
- 514 A big wet stain was on the round carpet.
- 513 Ripe pears are fit for a queen's table.
- 507 A bowl of rice is free with chicken stew.
- 515 The kite dipped and swayed, but stayed aloft.
- 518 This strong arm shall shield your honour.
- 504 The dusty bench stood by the stone wall.
- 519 She blushed when he gave her a white orchid.
- 520 The beetle droned in the hot June sun.

- 521 Press the pedal with your left foot.
- 523 The black trunk fell from the landing.
- 525 The theft of the pearl pin was kept secret.
- 522 Neat plans fail without luck.
- 534 Peep under the tent and see the clowns.
- 538 Flood the mails with requests for this book.
- 527 The vast space stretched into the far distance.
- 529 His wide grin earned many friends.
- 528 A rich farm is rare in this sandy waste.
- 530 Flax makes a fine brand of paper.
- 533 Even a just cause needs power to win.
- 537 A thing of small note can cause despair.
- 526 Shake hands with this friendly child.
- 524 The bank pressed for payment of the debt.
- 536 Cheap clothes are flashy but don't last.
- 531 Hurdle the pit with the aid of a long pole.
- 532 A strong bid may scare your partner stiff.
- 535 The leaf drifts along with a slow spin.

Appendix B

Statistical Analysis (Chapter 4)

ANOVA with repeated measures is a technique that is used to determine whether there are any statistically significant differences between the means of one or more independent groups that are based on repeated observations collected by different individuals. In our experiments, the independent groups are our speech conditions and the same participant listens to each of the five speech conditions. Therefore, the same people are being measured more than once on the same dependent variable which makes it a repeated measures design.

The repeated measures ANOVA tests for whether there are any differences between related participants means. The null hypothesis (H0) states that the means are equal:

$$H0 : \mu 1 = \mu 2 = \mu 3 = \ldots = \mu k \tag{B.1}$$

where μ = mean and k = number of speech conditions. The alternative hypothesis (HA) states that the related participants means are not equal:

$$HA : \textit{at least two means are significantly different} \tag{B.2}$$

ANOVA was applied for the WERs, self-reported cognitive load and naturalness scores. A p-value of less than 0.05 was used to reject the null hypothesis. In all cases the null hypothesis was rejected,therefore a post-hoc Tukey test with Bonferroni correction was used to determine which speech condition pairs are significantly different. The tables that follow show the significantly different pairs highlighted in green with the respective p-values for all experiments conducted in Chapter 4.

B.0.1 Significance results for Experiment 4.4.1: Semantically Unpredictable Sentences

Table 1: WER Pairs Exp. 1A

	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM		p=0.00693	p=0.01313	p=0.05166	p=0.02801
HMM			p=0.85062	p=0.10264	p=0.97212
Unit Selection				p=0.03281	p=0.66602
Hybrid					p=0.06932
Natural					

Figure B.1: SUS: 2011 Blizzard Dataset

Table 2: WER Pairs Exp. 1B

	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM		p=0.00537	p=0.09460	p=0.05536	p=0.07212
HMM			p=0.00012	p=0.80396	p=0.14655
Unit Selection				p=0.00072	p=0.00012
Hybrid					p=0.05371
Natural					

Figure B.2: SUS: 2010 Blizzard Dataset

Table 3: Self-reported Cognitive load Pairs for Exp. 1A

	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM		p=0.01466	p=0.14269	p=0.01466	p=0.00343
HMM			p=0.02441	p=0.03666	p=0.00355
Unit Selection				p=0.00982	p=0.02308
Hybrid					p=0.01466
Natural					

Figure B.3: SUS: 2011 Blizzard Dataset

Table 4: Self-reported Cognitive Load Pairs for Exp. 1B

	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM		p=0.18831	p=0.14252	p=0.02426	p=0.00282
HMM			p=0.01511	p=0.01453	p=0.00473
Unit Selection				p=0.00500	p=0.00142
Hybrid					p=0.18831
Natural					

Figure B.4: SUS: 2010 Blizzard Dataset

Table 5: Self-reported Naturalness Pairs for Exp. 1A

	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM		p=0.01154	p=0.00109	p=0.00116	p=0.00138
HMM			p=1.00000	p=0.03666	p=0.00228
Unit Selection				p=0.00476	p=0.00142
Hybrid					p=0.01154
Natural					

Figure B.5: SUS: 2011 Blizzard Dataset

Table 6: Self-reported Naturalness Pairs for Exp. 1B

	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM		p=0.12361	p=0.47729	p=0.00253	p=0.00153
HMM			p=0.50714	p=0.00473	p=0.00145
Unit Selection				p=0.00361	p=0.00144
Hybrid					p=0.12361
Natural					

Figure B.6: SUS: 2010 Blizzard Dataset

Table 7: GCA Parameter Estimate Significant Pairs for Exp. 1A

Intercept	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Linear	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Quadratic	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Cubic	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					

Figure B.7: SUS: 2011 Blizzard Dataset

Table 8: GCA Parameter Estimate Significant Pairs for Exp. 1B

Intercept	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Linear	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Quadratic	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Cubic	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					

Figure B.8: SUS: 2010 Blizzard Dataset

B.0.2 Significance results for Experiment 4.4.2: Semantically Meaningful Sentences

Table 9: WER Pairs Exp. 2

	LQ-HMM	HMM	Unit Selection	Natural	Hybrid
LQ-HMM		p=0.55085	p=0.42643	p=0.90009	p=0.03801
HMM			p=0.95466	p=0.37858	p=0.14671
Unit Selection				p=0.90009	p=0.05165
Natural					p=0.55085
Hybrid					

Figure B.9: SMS: 2011 Blizzard Dataset

Table 10: Self-reported Cognitive Load Pairs for Exp. 2

	LQ-HMM	HMM	Unit Selection	Natural	Hybrid
LQ-HMM		p=0.00514	p=0.03636	p=0.00064	p=0.00416
HMM			p=0.27547	p=0.00121	p=1.00000
Unit Selection				p=0.00052	p=0.30751
Natural					p=0.00514
Hybrid					

Figure B.10: SMS: 2011 Blizzard Dataset

Table 11: Self-reported Naturalness Pairs for Exp. 2

	LQ-HMM	HMM	Unit Selection	Natural	Hybrid
LQ-HMM		p=0.09727	p=0.02459	p=0.00046	p=0.00171
HMM			p=0.062702	p=0.00063	p=0.00775
Unit Selection				p=0.00063	p=0.04771
Natural					p=0.09727
Hybrid					

Figure B.11: SMS: 2011 Blizzard Dataset

Table 12: GCA Parameter Estimate Significant Pairs for Exp. 2

Intercept	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Linear	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Quadratic	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Cubic	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					

Figure B.12: SMS: 2011 Blizzard Dataset

B.0.3 Significance results for Experiment 4.4.3: Quiet vs Noise

Table 13: WER Pairs Exp. 3A

	LQ-HMM	HMM	Unit Selection	Natural	Hybrid
LQ-HMM		p=0.84692	p=0.00262	p=0.03830	p=0.55085
HMM			p=0.00836	p=0.04450	p=0.33026
Unit Selection				p=0.00072	p=0.05136
Natural					p=0.84692
Hybrid					

Figure B.13: -1dB: 2011 Blizzard Dataset

Table 14: WER Pairs Exp. 3B

	LQ-HMM	HMM	Unit Selection	Natural	Hybrid
LQ-HMM		p=0.71973	p=0.00153	p=0.38940	p=0.33612
HMM			p=0.00262	p=0.52448	p=0.13538
Unit Selection				p=0.00006	p=0.03015
Natural					p=0.71973
Hybrid					

Figure B.14: -3dB: 2011 Blizzard Dataset

Table 15: Self-reported Cognitive Load Pairs for Exp. 3A

	LQ-HMM	HMM	Unit Selection	Natural	Hybrid
LQ-HMM		p=0.45370	p=0.59408	p=0.01396	p=0.33074
HMM			p=0.03689	p=0.05562	p=0.91546
Unit Selection				p=0.00574	p=0.15808
Natural					p=0.45370
Hybrid					

Figure B.15: -1dB: 2011 Blizzard Dataset

Table 16: Self-reported Cognitive Load Pairs for Exp. 3B

	LQ-HMM	HMM	Unit Selection	Natural	Hybrid
LQ-HMM		p=0.82110	p=0.00781	p=0.38588	p=0.14387
HMM			p=0.01367	p=0.46444	p=0.18310
Unit Selection				p=0.06416	p=0.11375
Natural					p=0.82110
Hybrid					

Figure B.16: -3dB: 2011 Blizzard Dataset

Table 17: Self-reported Naturalness Pairs for Exp. 3A

	LQ-HMM	HMM	Unit Selection	Natural	Hybrid
LQ-HMM		p=0.01471	p=0.02341	p=0.00145	p=0.03666
HMM			p=0.80161	p=0.00834	p=1.00000
Unit Selection				p=0.00185	p=0.77681
Natural					p=0.01471
Hybrid					

Figure B.17: -1dB: 2011 Blizzard Dataset

Table 18: Self-reported Naturalness Pairs for Exp. 3B

	LQ-HMM	HMM	Unit Selection	Natural	Hybrid
LQ-HMM		p=0.07076	p=0.15808	p=0.00557	p=0.08041
HMM			p=1.00000	p=0.01038	p=0.77681
Unit Selection				p=0.01463	p=0.85141
Natural					p=0.07076
Hybrid					

Figure B.18: -3dB: 2011 Blizzard Dataset

Table 19: GCA Parameter Estimate Significant Pairs for Exp. 3A

Intercept	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Linear	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Quadratic	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Cubic	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					

Figure B.19: -1dB: 2011 Blizzard Dataset

Table 20: GCA Parameter Estimate Significant Pairs for Exp. 3B

Intercept	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Linear	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Quadratic	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					
Cubic	LQ-HMM	HMM	Unit Selection	Hybrid	Natural
LQ-HMM					
HMM					
Unit Selection					
Hybrid					
Natural					

Figure B.20: -3dB: 2011 Blizzard Dataset

Appendix C

Test Sentences used in Chapter 6

Semantically meaningful sentences taken from Glasgow herald newspaper used as test set for Chapter 6:

- 1078 The peace process will be buried!
- 1079 So it was at Easter Road.
- 1074 The world is split into two.
- 1093 It had been played at festivals.
- 1098 We may make a short-term appointment.
- 0026 They really shouldn't have put it out.
- 0042 A recording contract is on the horizon.
- 0036 He was released and was never indicted.
- 0043 Discussions with processors were needed over imports.
- 0032 And now that the pressure is off.
- 0031 Of course, this is nice to hear.
- 0046 I saw him about six months ago.
- 0052 The wingers must have felt the cold.
- 0028 This is not a floating Las Vegas.
- 1091 Then came the collision with Sergio.
- 0059 He was shocked by what he saw.
- 0037 There are some new buildings in there.
- 1092 Thirteen people will be made redundant.

- 0045 Just when you thought it was safe.
- 1088 I see no need for change.
- 0062 It comes on like a tidal wave.
- 0063 It's an amazing sight, isn't it ?
- 0065 We welcome the decision of Mr Byers.
- 0081 Next league matches, Aberdeen - Dundee United.
- 0080 Mummy believed in one thing above all.
- 0096 He tells the truth and that's important.
- 1007 Meanwhile, the business of government goes on.
- 0067 Meanwhile, the weather was causing widespread disruption.
- 0098 What happens when the funding is exhausted.
- 1017 A judicial review is also being considered.
- 0092 For once, you should believe the hype.
- 0078 This means an awful lot to me.
- 0076 If convicted, they face the death penalty.
- 0093 Why did they let him go ?
- 0085 We are making progress on waiting lists.
- 1009 That figure will always remain with me.
- 0099 He then had a swipe at Hearts.
- 0089 Then it was the next big thing.
- 1005 You have to rely on each other.
- 0074 I've never worked in a bar before.
- 1013 They hold on for many years.
- 1020 We have never been short of volunteers.
- 1024 I don't know why you say goodbye.
- 1087 We have taken steps to rectify it.

- 1084 We are now in an election period.
- 1071 Three years probation is just a joke.
- 1081 The dictator of Iraq is not disarming.
- 1047 However, there is another aspect of this.
- 1062 He also received a medal of honour.
- 1072 Inclusion and Autism, Is It Working ?
- 1040 I had a sense of deja vu.
- 1034 It is a form of physical exercise.
- 1060 This meeting is the path to salvation.
- 1082 It seemed a moving and fitting tribute.
- 1069 I think there's not much between them.
- 1083 This is the logic of punishment assaults.
- 1094 No finals are to be contested today.
- 1068 This could be a recipe for conflict.
- 1037 Every school in Scotland should be excellent.
- 1073 Could you put it in writing ?
- 1095 This is not a struggle against Islam.
- 1099 Among them was Gary Robertson from Dundee.
- 0010 People look, but no one ever finds it.
- 1006 There is no margin for error.
- 0090 Why was he in such a hurry ?
- 0073 Can you imagine a world without design ?
- 0066 This, he told the committee, was not forthcoming.
- 0017 Others have tried to explain the phenomenon physically.
- 1080 This is not, however, a General Election reshuffle.
- 1089 Therefore, this type of aircraft is completely safe.

- 1086 However, the move was bitterly criticised last night.
- 1067 It's a matter of huge concern.
- 0049 For the meantime, though, the signs are good.
- 0058 What on earth is wrong with you ?
- 1038 I could not have run at all, otherwise.
- 1064 I am a soldier, a soldier, a soldier.
- 0040 However, details have yet to be worked out.
- 1031 Times have changed, but have they improved ?
- 1090 As a nation, we must become more active.
- 1036 Hopefully, it will be built by next year.
- 0020 Many complicated ideas about the rainbow have been formed.
- 0030 He probably would not have added, mainly by me.
- 0053 In time, the First Minister will grow in stature.
- 1066 They were going through the motions, but that's about all.
- 1050 I'm back playing football, which is what it's all about.
- 1085 He served in the Gulf, the Falklands, and Northern Ireland.
- 0082 For me, the tour is wide open this year.
- 1053 And also, whatever he does, can he really win ?
- 1065 We always thought it should be digital, and cheap.
- 0013 Some have accepted it as a miracle without physical explanation.
- 0100 According to the criteria, he is qualified for Scotland.
- 0064 Mr Cook, a left-winger from Britain, would be ideally placed.
- 0002 Ask her to bring these things with her from the store.
- 1012 This is a one-year deal, but who knows ?
- 1033 People don't feel safe, including me.
- 1045 So we said, no, we're not going to do that.

- 1001 We are a party for people, not against people.
- 0083 She opens tonight in Glasgow, at the King's Theatre.
- 1027 They make it look very easy.
- 0025 We have to pull together, or we will hang apart.
- 0012 Throughout the centuries people have explained the rainbow in various ways.
- 1015 It may mean more money, but we don't need the money.
- 1019 It used to bother me sometimes, but it doesn't any more.
- 0007 The rainbow is a division of white light into many beautiful colors.
- 1100 They were no longer together, he said, and her life was in tatters.
- 1000 Read the book and see the film, or the other way round ?
- 1061 Of course we make mistakes, but we don't make too many.
- 0072 Chris Smith, the culture secretary, said he was satisfied with the decision.
- 1077 He always has been, since his early days with Scottish Opera.
- 1049 If they don't understand official documents, they take them along to us.
- 0084 He had been drinking the night before, but was still over the limit.
- 1056 The prime minister, also unaware of the e-mail, was drawn in last night.
- 0033 Neither is likely to win the argument, because too many questions remain unanswered.
- 1023 That is partly true, but there are ways of doing it.
- 0050 Of the threat to scrap the sleepers, he said, It would be madness.
- 1055 I am pleased with the result, as it was a possible upset.
- 0039 In the past three years or so, everything has changed for the orchestra.
- 1054 Alan Main, the keeper, however, has been in a similar position before.
- 1051 Ritual resentment is nothing new in football, or any sport for that matter.
- 0097 It is one of those things, but it is frustrating for us.

Appendix D

Statistical Analysis (Chapter 6)

ANOVA with repeated measures is a technique that is used to determine whether there are any statistically significant differences between the means of one or more independent groups that are based on repeated observations collected by different individuals. In our experiments, the independent groups are our speech conditions and the same participant listens to each of the five speech conditions. Therefore, the same people are being measured more than once on the same dependent variable which makes it a repeated measures design.

The repeated measures ANOVA tests for whether there are any differences between related participants means. The null hypothesis (H0) states that the means are equal:

$$H0 : \mu 1 = \mu 2 = \mu 3 = \ldots = \mu k \tag{D.1}$$

where μ = mean and k = number of speech conditions. The alternative hypothesis (HA) states that the related participants means are not equal:

$$HA : \textit{at least two means are significantly different} \tag{D.2}$$

ANOVA was applied for the WERs, self-reported cognitive load and naturalness scores. A p-value of less than 0.05 was used to reject the null hypothesis. In all cases the null hypothesis was rejected,therefore a post-hoc Tukey test with Bonferroni correction was used to determine which speech condition pairs are significantly different. The tables that follow show the significantly different pairs highlighted in green with the respective p-values for all experiments conducted in Chapter 6.

Table 1: WER Pairs Exp. 1

	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)		p=0.90897	p=0.10448	p=0.89957	p=0.69675	p=0.08294
B (Vocoded)			p=0.31679	p=0.88640	p=0.94305	p=0.10738
C (Predicted F0)				p=0.09738	p=0.50339	p=0.95443
D (Predicted MCC)					p=0.40284	p=0.10692
E (Natural Duration)						p=0.36956
F(full DNN)						

Figure D.1: Quiet: Nick harvard

Table 2: WER Pairs Exp. 2A

	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)		p=0.20120	p=0.34655	p=0.14152	p=0.00014	p=0.00005
B (Vocoded)			p=0.71122	p=0.00245	p=0.00053	p=0.00002
C (Predicted F0)				p=0.00067	p=0.00004	p=0.00001
D (Predicted MCC)					p=0.02685	p=0.00004
E (Natural Duration)						p=0.36922
F(full DNN)						

Figure D.2: -1dB: Nick harvard

Table 3: WER Pairs Exp. 2B

	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)		p=0.31292	p=0.19563	p=0.00000	p=0.00004	p=0.00021
B (Vocoded)			p=1.00000	p=0.00003	p=0.00001	p=0.00001
C (Predicted F0)				p=0.00042	p=0.00010	p=0.00004
D (Predicted MCC)					p=0.36042	p=0.14469
E (Natural Duration)						p=0.31241
F(full DNN)						

Figure D.3: -3dB: Nick harvard

Table 4: WER Pairs Exp. 2C

	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)		p=0.35246	p=0.02893	p=0.00169	p=0.00004	p=0.00000
B (Vocoded)			p=0.28619	p=0.00141	p=0.00052	p=0.00004
C (Predicted F0)				p=0.00023	p=0.00000	p=0.00000
D (Predicted MCC)					p=0.54605	p=0.00714
E (Natural Duration)						p=0.11338
F(full DNN)						

Figure D.4: -5dB: Nick harvard

Table 5: Self-reported Cognitive load Pairs for Exp. 1

	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)		p=0.05633	p=0.03525	p=0.00269	p=0.00139	p=0.00038
B (Vocoded)			p=0.73957	p=0.10088	p=0.06833	p=0.01280
C (Predicted F0)				p=0.12183	p=0.29092	p=0.04079
D (Predicted MCC)					p=1.00000	p=0.69497
E (Natural Duration)						p=0.78304
F(full DNN)						

Figure D.5: Quiet: Nick harvard

Table 6: Self-reported Cognitive load Pairs for Exp. 2A

	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)		p=0.25971	p=0.25536	p=0.04371	p=0.01409	p=0.00223
B (Vocoded)			p=0.62702	p=0.41693	p=0.11691	p=0.01628
C (Predicted F0)				p=0.73983	p=0.23058	p=0.06260
D (Predicted MCC)					p=0.27593	p=0.13079
E (Natural Duration)						p=0.11995
F(full DNN)						

Figure D.6: -1dB: Nick harvard

Table 7: Self-reported Cognitive load Pairs for Exp. 2B

	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)		p=0.85141	p=0.85100	p=0.01151	p=0.11313	p=0.09629
B (Vocoded)			p=0.56532	p=0.00122	p=0.04416	p=0.07707
C (Predicted F0)				p=0.00918	p=0.08964	p=0.30883
D (Predicted MCC)					p=0.30751	p=0.19655
E (Natural Duration)						p=0.91436
F(full DNN)						

Figure D.7: -3dB: Nick harvard

Table 8: Self-reported Cognitive load Pairs for Exp. 2C

	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)		p=0.40702	p=0.45746	p=0.01503	p=0.03545	p=0.01357
B (Vocoded)			p=0.91634	p=0.09694	p=0.16159	p=0.02459
C (Predicted F0)				p=0.13726	p=0.09727	p=0.03301
D (Predicted MCC)					p=1.00000	p=0.62702
E (Natural Duration)						p=0.12705
F(full DNN)						

Figure D.8: -5dB: Nick harvard

Table 9: Self-reported Naturalness Pairs for Exp. 1

	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)		p=0.06311	p=0.00436	p=0.00041	p=0.00010	p=0.00027
B (Vocoded)			p=0.65968	p=0.16256	p=0.03580	p=0.01972
C (Predicted F0)				p=0.10429	p=0.01166	p=0.02232
D (Predicted MCC)					p=0.28604	p=0.23878
E (Natural Duration)						p=1.00000
F(full DNN)						

Figure D.9: Quiet: Nick harvard

Table 10: Self-reported Naturalness Pairs for Exp. 2A

	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)		p=0.77681	p=0.48402	p=0.02341	p=0.02090	p=0.03301
B (Vocoded)			p=0.77681	p=0.14459	p=0.03301	p=0.04982
C (Predicted F0)				p=0.14503	p=0.03666	p=0.07767
D (Predicted MCC)					p=0.53795	p=0.66554
E (Natural Duration)						p=0.78973
F(full DNN)						

Figure D.10: -1dB: Nick harvard

Table 11: Self-reported Naturalness Pairs for Exp. 2B

	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)		p=0.09534	p=0.49219	p=0.06021	p=0.09182	p=0.00352
B (Vocoded)			p=0.47729	p=0.34049	p=0.82964	p=0.11995
C (Predicted F0)				p=0.09727	p=0.18166	p=0.03301
D (Predicted MCC)					p=0.52419	p=0.67451
E (Natural Duration)						p=0.11981
F(full DNN)						

Figure D.11: -3dB: Nick harvard

Table 12: Self-reported Naturalness Pairs for Exp. 2C

	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)		p=0.39295	p=0.06560	p=0.24017	p=0.08071	p=0.04982
B (Vocoded)			p=0.10960	p=0.77681	p=0.30529	p=0.15209
C (Predicted F0)				p=0.17357	p=1.00000	p=1.00000
D (Predicted MCC)					p=0.24017	p=0.24017
E (Natural Duration)						p=1.00000
F(full DNN)						

Figure D.12: -5dB: Nick harvard

Table 13: GCA Parameter Estimate Significant Pairs for Exp. 1

Intercept	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Linear	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Quadratic	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Cubic	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Figure D.13: Quiet: Nick harvard

Table 14: GCA Parameter Estimate Significant Pairs for Exp. 2A

Intercept	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Linear	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Quadratic	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Cubic	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Figure D.14: -1dB: Nick harvard

Table 15: GCA Parameter Estimate Significant Pairs for Exp. 2B

Intercept	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Linear	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Quadratic	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Cubic	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Figure D.15: -3dB: Nick harvard

Table 16: GCA Parameter Estimate Significant Pairs for Exp. 2C

Intercept	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Linear	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Quadratic	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Cubic	A (Natural)	B (Vocoded)	C (Predicted F0)	D (Predicted MCC)	E (Natural Duration)	F(full DNN)
A (Natural)						
B (Vocoded)						
C (Predicted F0)						
D (Predicted MCC)						
E (Natural Duration)						
F(full DNN)						

Figure D.16: -5dB: Nick harvard

Appendix E

Test Sentences used in Chapter 7

- 0039 The Roman type of all these printers is similar in character,
- 0009 On the 27th April, in the following year,
- 0172 For this he was summoned before a magistrate, and sentenced as already stated.
- 0272 on the opening of Whitecross prison for debtors in 1815.
- 0034 This gradually was forced upon the consciousness of the Corporation,
- 0065 All the misdemeanants, whatever their offense, were lodged in this chapel ward.
- 0108 Spirits were freely introduced, and although he at first abstained,
- 0113 A prisoner, generally the oldest and most dexterous thief,
- 0118 Various punishments were inflicted, the heaviest of which was standing in the pillory.
- 0135 She went up to the ward and found him lying down, quote,
- 0182 The tried and the untried, young and old, were herded together
- 0309 Proper hours for locking and unlocking prisoners should be insisted upon;
- 0129 Many of the jails were in the most deplorable condition:
- 0099 and report at length upon the condition of the prisons of the country.
- 0105 It was long before the many jurisdictions imitated the few.
- 0100 many without stockings, and with hardly shoes to their feet;
- 0196 the provision of dining-rooms and dining-tables.
- 0112 He was death's counterfeit, tall, shriveled, and pale;
- 0317 The fatal news was not always received in the same way.

- 0015 I will quote an extract from the reverend gentleman's own journal.
- 0231 Although these objectionable practices had disappeared,
- 0138 and died in a manner strongly contrasting with that of his fellows.
- 0018 The crime, long carried on without detection, was first discovered in 1820,
- 0020 had been sold out under a forged power of attorney.
- 0129 learnt through the firm's correspondence that a quantity of gold-dust
- 0185 raised the alarm, and suspicion fell upon the three murderers, who were arrested.
- 0081 Banks and bankers continued to be victimized.
- 0187 near it was his Waterloo medal, and the above-mentioned ten-pound note.
- 0112 A telegraphic message, then newly adapted to the purposes of criminal detection,
- 0240 without capital, and at railroad speed.
- 0252 When he knew that he could not escape his fate,
- 0393 called for a roast duck directly he entered the condemned cell.
- 0217 Navigation and discipline could not be easy with such a nondescript crew.
- 0231 namely, to suppress it and substitute another.
- 0084 The greatest pains might be taken to secure isolation,
- 0269 More animal food was given than was necessary.
- 0295 A single officer was the only custodian and disciplinary authority in the jail.
- 0015 Beat hard one minute before pouring in the yeast.
- 0090 Active exercise in like circumstances tempts debility and disease.
- 0042 In this we have had assistance from many bankers and businessmen,
- 0145 From those willing to join in establishing this hoped-for period of peace,
- 0161 just as other countries have had them for over a decade.
- 0176 Did England let nature take her course? No.
- 0014 the creation of a useful instrument for man ultimately comes.
- 0021 or to one industry, or to an individual private occupation.

- 0147 we must continue to protect children,
- 0154 I consider this legislation a positive recovery measure.
- 0187 but these twenty years have shown by experience definite possibilities for improvement.
- 0203 For that we can be thankful to the God who watches over America.
- 0005 I take this means of saying thank you.
- 0025 National laws are needed to complete that program.
- 0019 And there may be only nine.
- 0024 Why was the age fixed at seventy?
- 0029 There is general approval so far as the lower federal courts are concerned.
- 0097 Cellulose is widespread as a constituent of the skeletons of the lower animals
- 0137 but it suggests a suspicion of their identity which needs careful testing.
- 0081 exhibit a remarkable series of homologies pointing to a five-toed ancestor,
- 0018 Like a great ball of fire the sun sinks in the west.
- 0155 Lumps of bitumen are found in great abundance in this river.
- 0166 The center of each division of the town is occupied by a fortress.
- 0199 there suddenly appeared upon the wall an armless hand.
- 0331 After this second exploit, his praise was in all mouths.
- 0500 The ruins reach the height of about forty feet.
- 0179 She watched as he slumped down with an empty expression on his face.
- 0033 Then he experienced his first sensation of pain, which became excruciating.
- 0034 The Governor was lifted onto a stretcher and taken into trauma room 2.
- 0142 The Vice President conferred with White House Assistant Press Secretary Malcolm Kilduff
- 0162 other terminal buildings and the neighboring parking lots, of all people.
- 0024 After searching their records from 10 p.m. to 4 a.m.
- 0051 On that date, Klein's placed an internal control number VC836 on this rifle.
- 0084 It was more bulky toward the bottom, end quote, than toward the top.

    - 0031 The box on the floor, behind the three near the window,
    - 0061 The results of this investigation are fully discussed in chapter 6, page 249.
    - 0017 He heard two more shots spaced, quote, pretty well even to me.
    - 0151 He was walking into the office from the back hallway,
    - 0086 Her reaction when she saw Oswald in the lineup was that, quote,
    - 0103 Two other important eyewitnesses to Oswald's flight were Ted Callaway,
    - 0206 The balance due on the purchase was $19.95.0232 As has been discussed previously$,
- 0240 At 1:24 p.m., the police radio reported, quote,
- 0006 police sirens sounded along Jefferson Boulevard.
- 0128 He also denied that he had received the rifle through this box.
- 0151 He acknowledged the encounter with the police officer on the second floor.
- 0226 Oswald had apparently mistaken the county jail for the city jail.
- 0152 It will be recalled from the discussion in chapter 3
- 0211 This would make him approximately 10, well, almost 11 years old. End quote.
- 0178 not unpopular, end quote, at that time. Donovan testified
- 0196 Oswald's decided rejection of both capitalism and communism
- 0234 quote, except in the US, the living standard is a little higher.
- 0003 Background and Possible Motives, Part 4.
- 0074 where he expressed a reluctance to work in the industrial field.
- 0158 As for my return entrance visa please consider it separately. End quote.
- 0012 there have been attempts on the lives of one out of every three.
- 0032 or their desire to have frequent and easy access to the people.
- 0145 The Secret Service responded, quote,
- 0251 to the other agencies. No specific guidance was provided.
- 0022 provided by other agencies.
- 0084 and his application was approved on the following day.

- 0193 and so advised the Dallas office in the ordinary course of business.
- 0026 He gave evidence of settling down.
- 0040 The Commission appreciates the large volume of cases handled by the FBI
- 0068 as requiring evidence of a plan or conspiracy to injure the President.
- 0095 The Service prefers to have two agents perform advance preparations.
- 0139 The Service had 28 agents participating in the Dallas visit.
- 0200 paying particular attention to the crowd for any unusual activity.

Appendix F

Statistical Analysis (Chapter 7)

ANOVA with repeated measures is a technique that is used to determine whether there are any statistically significant differences between the means of one or more independent groups that are based on repeated observations collected by different individuals. In our experiments, the independent groups are our speech conditions and the same participant listens to each of the five speech conditions. Therefore, the same people are being measured more than once on the same dependent variable which makes it a repeated measures design.

The repeated measures ANOVA tests for whether there are any differences between related participants means. The null hypothesis (H0) states that the means are equal:

$$H0 : \mu 1 = \mu 2 = \mu 3 = \ldots = \mu k \tag{F.1}$$

where μ = mean and k = number of speech conditions. The alternative hypothesis (HA) states that the related participants means are not equal:

$$HA : \textit{at least two means are significantly different} \tag{F.2}$$

ANOVA was applied for the WERs, self-reported cognitive load and naturalness scores. A p-value of less than 0.05 was used to reject the null hypothesis. In all cases the null hypothesis was rejected,therefore a post-hoc Tukey test with Bonferroni correction was used to determine which speech condition pairs are significantly different. The tables that follow show the significantly different pairs highlighted in green with the respective p-values for all experiments conducted in Chapter 7.

F.0.1 Significance results for Experiment 7.4.1: 2020 state-of-the-art models

Table 1: WER Pairs Exp. 1A

	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human		p=0.24675	p=0.00002	p=0.23447	p=0.00074
Tacotron 2			p=0.00146	p=0.03267	p=0.03013
DC-TTS				p=0.00007	p=0.00745
WaveRNN					p=0.00146
Unit Selection					

Figure F.1: -3dB: LJSpeech

Table 2: WER Pairs Exp. 1B

	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human		p=0.27543	p=0.02543	p=0.23447	p=0.00654
Tacotron 2			p=0.01543	p=0.04277	p=0.30013
DC-TTS				p=0.00326	p=0.08312
WaveRNN					p=0.00254
Unit Selection					

Figure F.2: -5dB: LJSpeech

Table 3: Self-reported Cognitive load Pairs for Exp. 1A

	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human		p=0.11510	p=0.00039	p=0.64441	p=0.00314
Tacotron 2			p=0.00006	p=0.14459	p=0.05719
DC-TTS				p=0.00015	p=0.00516
WaveRNN					p=0.00182
Unit Selection					

Figure F.3: -3dB: LJSpeech

Table 4: Self-reported Cognitive load Pairs for Exp. 1B

	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human		p=0.34578	p=0.03201	p=0.34578	p=0.03689
Tacotron 2			p=0.08897	p=1.00000	p=0.14891
DC-TTS				p=0.02627	p=0.17357
WaveRNN					p=0.23304
Unit Selection					

Figure F.4: -5dB: LJSpeech

Table 5: Self-reported Naturalness Pairs for Exp. 1A

	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human		p=0.22540	p=0.00463	p=0.46444	p=0.00282
Tacotron 2			p=0.01334	p=0.28180	p=0.02712
DC-TTS				p=0.00629	p=0.56417
WaveRNN					p=0.00359
Unit Selection					

Figure F.5: -3dB: LJSpeech

Table 6: Self-reported Naturalness Pairs for Exp. 1B

	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human		p=0.14891	p=0.17357	p=0.42371	p=0.34470
Tacotron 2			p=1.00000	p=0.77283	p=1.00000
DC-TTS				p=0.34578	p=0.76559
WaveRNN					p=0.77283
Unit Selection					

Figure F.6: -5dB: LJSpeech

Table 7: GCA Parameter Estimate Significant Pairs for Exp. 1A

Intercept	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human					
Tacotron 2					
DC-TTS					
WaveRNN					
Unit Selection					
Linear	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human					
Tacotron 2					
DC-TTS					
WaveRNN					
Unit Selection					
Quadratic	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human					
Tacotron 2					
DC-TTS					
WaveRNN					
Unit Selection					
Cubic	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human					
Tacotron 2					
DC-TTS					
WaveRNN					
Unit Selection					

Figure F.7: -3dB: LJSpeech

Table 8: GCA Parameter Estimate Significant Pairs for Exp. 1B

Intercept	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human					
Tacotron 2					
DC-TTS					
WaveRNN					
Unit Selection					
Linear	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human					
Tacotron 2					
DC-TTS					
WaveRNN					
Unit Selection					
Quadratic	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human					
Tacotron 2					
DC-TTS					
WaveRNN					
Unit Selection					
Cubic	Human	Tacotron 2	DC-TTS	WaveRNN	Unit Selection
Human					
Tacotron 2					
DC-TTS					
WaveRNN					
Unit Selection					

Figure F.8: -5dB: LJSpeech

F.0.2 Significance results for Experiment 7.4.2: 2022 state-of-the-art models

Table 9: WER Pairs Exp. 2A

	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human		p=0.26543	p=0.20876	p=0.25234	p=0.00564
Multiband-melGAN			p=0.04531	p=0.02187	p=0.02432
Tacotron 2				p=0.01325	p=0.00564
Fastspeech 2					p=0.00213
Merlin					

Figure F.9: -1dB: LJSpeech

Table 10: WER Pairs Exp. 2B

	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human		p=0.38761	p=0.27861	p=0.01665	p=0.03112
Multiband-melGAN			p=0.00322	p=0.00845	p=0.03467
Tacotron 2				p=0.02186	p=0.04327
Fastspeech 2					p=0.17543
Merlin					

Figure F.10: -3dB: LJSpeech

Table 11: WER Pairs Exp. 2C

	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human		p=0.11432	p=0.14622	p=0.26781	p=0.19762
Multiband-melGAN			p=0.29645	p=0.65372	p=0.34701
Tacotron 2				p=0.38753	p=0.25424
Fastspeech 2					p=0.03412
Merlin					

Figure F.11: -5dB: LJSpeech

Table 12: Self-reported Cognitive load Pairs for Exp. 2A

	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human		p=0.43741	p=0.74459	p=0.06021	p=0.00090
Multiband-melGAN			p=0.23878	p=0.08773	p=0.00203
Tacotron 2				p=0.09396	p=0.03410
Fastspeech 2					p=0.16579
Merlin					

Figure F.12: -1dB: LJSpeech

Table 13: Self-reported Cognitive load Pairs for Exp. 2B

	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human		p=0.03301	p=0.58775	p=0.00124	p=0.00462
Multiband-melGAN			p=0.09727	p=0.00049	p=0.00132
Tacotron 2				p=0.00084	p=0.00192
Fastspeech 2					p=1.00000
Merlin					

Figure F.13: -3dB: LJSpeech

Table 14: Self-reported Cognitive load Pairs for Exp. 2C

	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human		p=0.10375	p=0.01974	p=0.49592	p=0.14881
Multiband-melGAN			p=0.67451	p=0.00466	p=0.02622
Tacotron 2				p=0.00136	p=0.00436
Fastspeech 2					p=0.59992
Merlin					

Figure F.14: -5dB: LJSpeech

Table 15: Self-reported Naturalness Pairs for Exp. 2A

	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human		p=0.85141	p=0.10721	p=0.01028	p=0.00064
Multiband-melGAN			p=0.16005	p=0.02459	p=0.00292
Tacotron 2				p=0.43741	p=0.01038
Fastspeech 2					p=0.02475
Merlin					

Figure F.15: -1dB: LJSpeech

Table 16: Self-reported Naturalness Pairs for Exp. 2B

	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human		p=0.23878	p=0.23878	p=0.81285	p=0.04931
Multiband-melGAN			p=0.02090	p=0.22297	p=0.00852
Tacotron 2				p=0.52304	p=0.11133
Fastspeech 2					p=0.03565
Merlin					

Figure F.16: -3dB: LJSpeech

Table 17: Self-reported Naturalness Pairs for Exp. 2C

	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human		p=0.23878	p=0.23878	p=0.81285	p=0.04931
Multiband-melGAN			p=0.02090	p=0.22297	p=0.00852
Tacotron 2				p=0.52304	p=0.11133
Fastspeech 2					p=0.03565
Merlin					

Figure F.17: -5dB: LJSpeech

Table 18: GCA Parameter Estimate Significant Pairs for Exp. 2A

Intercept	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human					
Multiband-melGAN					
Tacotron 2					
Fastspeech 2					
Merlin					
Linear	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human					
Multiband-melGAN					
Tacotron 2					
Fastspeech 2					
Merlin					
Quadratic	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human					
Multiband-melGAN					
Tacotron 2					
Fastspeech 2					
Merlin					
Cubic	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human					
Multiband-melGAN					
Tacotron 2					
Fastspeech 2					
Merlin					

Figure F.18: -1dB: LJSpeech

Table 19: GCA Parameter Estimate Significant Pairs for Exp. 2B

Intercept	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human					
Multiband-melGAN					
Tacotron 2					
Fastspeech 2					
Merlin					

Linear	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human					
Multiband-melGAN					
Tacotron 2					
Fastspeech 2					
Merlin					

Quadratic	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human					
Multiband-melGAN					
Tacotron 2					
Fastspeech 2					
Merlin					

Cubic	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human					
Multiband-melGAN					
Tacotron 2					
Fastspeech 2					
Merlin					

Figure F.19: -3dB: LJSpeech

Table 20: GCA Parameter Estimate Significant Pairs for Exp. 2C

Intercept	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human					
Multiband-melGAN					
Tacotron 2					
Fastspeech 2					
Merlin					
Linear	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human					
Multiband-melGAN					
Tacotron 2					
Fastspeech 2					
Merlin					
Quadratic	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human					
Multiband-melGAN					
Tacotron 2					
Fastspeech 2					
Merlin					
Cubic	Human	Multiband-melGAN	Tacotron 2	Fastspeech 2	Merlin
Human					
Multiband-melGAN					
Tacotron 2					
Fastspeech 2					
Merlin					

Figure F.20: -5dB: LJSpeech

Bibliography

Sylvia K Ahern. *Activation and intelligence: Pupillometric correlates of individual differences in cognitive abilities.* PhD thesis, ProQuest Information & Learning, 1978.

Sara Alhanbali, Piers Dawes, Simon Lloyd, and Kevin J Munro. Self-reported listening-related effort and fatigue in hearing-impaired adults. *Ear and Hearing*, 38(1):e39–e48, 2017.

Shota Amada, Ryosuke Sugiura, Yutaka Kamamoto, Noboru Harada, Takehiro Moriya, Takeshi Yamada, and Shoji Makino. Experimental evaluation of wavernn predictor for audio lossless coding. In *The Acoustical Society of Japan 1018Autumn Meeting*, pages 1149–1152, 2018.

Hasan Ayaz, Paul Crawford, Adrian Curtin, Mashaal Syed, Banu Onaral, Willem M Beltman, and Patricia A Shewokis. Differential prefrontal response during natural and synthetic speech perception: An fnir based neuroergonomics study. In *International Conference on Augmented Cognition*, pages 241–249. Springer, 2013.

Jackson Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin*, 91(2):276, 1982.

Jackson Beatty, Brennis Lucero-Wagoner, et al. The pupillary system. *Handbook of psychophysiology*, 2(142-162), 2000.

Christian Benoît, Martine Grice, and Valérie Hazan. The sus test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication*, 18(4):381–392, 1996.

Randall R Benson, DH Whalen, Matthew Richardson, Brook Swainson, Vincent P Clark, Song Lai, and Alvin M Liberman. Parametrically dissociating speech and nonspeech perception in the brain using fMRI. *Brain and language*, 78(3):364–396, 2001.

Corinna Bernarding, Daniel J Strauss, Ronny Hannemann, and Farah I Corona-Strauss. Quantification of listening effort correlates in the oscillatory eeg activity: a feasibility study. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4615–4618. IEEE, 2012.

Mark Beutnagel, Mehryar Mohri, and Michael Riley. Rapid unit selection from a large speech corpus for concatenative speech synthesis. In *Sixth European Conference on Speech Communication and Technology*, 1999.

Alan Black, Paul Taylor, Richard Caley, and Rob Clark. The festival speech synthesis system, 1998.

Alan W Black and Paul A Taylor. Automatically clustering similar units for unit selection in speech synthesis. 1997.

Alan W Black, Heiga Zen, and Keiichi Tokuda. Statistical parametric speech synthesis. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–1229. IEEE, 2007.

Carol Chermaz, Cassia Valentini-Botinhao, Henning F Schepker, and Simon King. Evaluating near end listening enhancement algorithms in realistic environments. In *INTERSPEECH*, pages 1373–1377, 2019.

Alistair Conkie. Robust unit selection system for speech synthesis. In *137th meeting of the Acoustical Society of America*, page 978, 1999.

Martin Cooke, Catherine Mayo, Cassia Valentini-Botinhao, Yannis Stylianou, Bastian Sauert, and Yan Tang. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55(4):572–585, 2013.

Angela Damian, Farah I Corona-Strauss, Ronny Hannemann, and Daniel J Strauss. Towards the assessment of listening effort in real life situations: Mobile eeg recordings in a multimodal driving situation. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 8123–8126. IEEE, 2015.

Cristina Delogu, Stella Conte, and Ciro Sementina. Cognitive factors in the evaluation of synthetic speech. *Speech Communication*, 24(2):153–168, 1998.

Jamie L Desjardins and Karen A Doherty. Age-related changes in listening effort for various types of masker noises. *Ear and hearing*, 34(3):261–272, 2013.

Jamie L Desjardins and Karen A Doherty. The effect of hearing aid noise reduction on listening effort in hearing-impaired adults. *Ear and Hearing*, 35(6):600–610, 2014.

Susan A Duffy and David B Pisoni. Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*, 35(4):351–389, 1992.

Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Fatchord. Wavernn. `https://github.com/fatchord/WaveRNN.git`, 2019.

Sarah Fraser, Jean-Pierre Gagné, Majolaine Alepins, and Pascale Dubois. Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of speech, language, and hearing research*, 2010.

Toshiaki Fukada, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. An adaptive algorithm for mel-cepstral analysis of speech. In *Proc. ICASSP*, volume 1, pages 137–140, 1992.

Jean-Pierre Gagne, Jana Besser, and Ulrike Lemke. Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends in hearing*, 21:1–25, 2017.

Penny Anderson Gosselin and Jean-Pierre Gagné. Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research*, 2011.

Avashna Govender. Merlyn. https://github.com/AvashnaGovender/Merlyn, 2019.

Avashna Govender and Simon King. Measuring the cognitive load of synthetic speech using a dual task paradigm. In *Interspeech*, pages 2843–2847, 2018a.

Avashna Govender and Simon King. Using pupillometry to measure the cognitive load of synthetic speech. *System*, 50:100, 2018b.

Avashna Govender, Cassia Valentini-Botinhao, and Simon King. Measuring the contribution to cognitive load of each predicted vocoder speech parameter in dnn-based speech synthesis. In *Submitted to Speech Synthesis Workshop (SSW)*, volume 2019, 2019a.

Avashna Govender, Anita E Wagner, and Simon King. Using pupil dilation to measure cognitive load when listening to text-to-speech in quiet and in noise. *Proc. Interspeech 2019*, pages 1551–1555, 2019b.

Eren Gölge. Ddc tacotron2. https://colab.research.google.com/drive/1tKHSI20kRlOL0PSA8mCVJQIrgRIswg0F 2020.

Karen S Helfer, Jamie Chevalier, and Richard L Freyman. Aging, spatial cues, and single-versus dual-task performance in competing speech perception. *The Journal of the Acoustical Society of America*, 128(6):3625–3633, 2010.

Eckhard H Hess and James M Polt. Pupil size in relation to mental activity during simple problem-solving. *Science*, 143(3611):1190–1192, 1964.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Benjamin WY Hornsby. The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear and hearing*, 34(5):523–534, 2013.

Andrew J Hunt and Alan 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*, volume 1, pages 373–376. IEEE, 1996.

Keith Ito and Linda Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

Ingrid S Johnsrude and Jennifer M Rodd. Factors that increase processing demands when listening to speech. In *Neurobiology of Language*, pages 491–502. Elsevier, 2016.

S Joshi, RM Kalwani, and JI Gold. The relationship between locus coeruleus neuronal activity and pupil diameter. In *Society for Neuroscience Abstracts*, 2013.

Daniel Kahneman and Jackson Beatty. Pupil diameter and load on memory. *Science*, 154(3756): 1583–1585, 1966.

Daniel Kahnemann and Jackson Beatty. Pupillary responses in a pitch-discrimination task. *Attention, Perception, & Psychophysics*, 2(3):101–105, 1967.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. *arXiv preprint arXiv:1802.08435*, 2018.

Hideki Kawahara. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6):349–353, 2006.

Sangramsing Kayte, Monica Mundada, and Charansing Kayte. A review of unit selection speech synthesis. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(5), 2015.

Simon King. An introduction to statistical parametric speech synthesis. *Sadhana*, 36(5):837–852, 2011.

Simon King and Vasilis Karaiskos. The Blizzard Challenge 2010. Sept. 2010.

Simon King and Vasilis Karaiskos. The Blizzard Challenge 2011. Italy, Sept. 2011.

Simon King, Lovisa Wihlborg, and Wei Guo. The Blizzard Challenge 2017. Stockholm, Sept. 2017.

Jeff Klingner, Barbara Tversky, and Pat Hanrahan. Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, 48(3):323–332, 2011.

Thomas Koelewijn, Adriana A Zekveld, Joost M Festen, and Sophia E Kramer. Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, 33(2): 291–300, 2012.

Sophia E Kramer, Artur Lorens, Frans Coninx, Adriana A Zekveld, Anna Piotrowska, and Henryk Skarzynski. Processing load during listening: The influence of task characteristics on the pupil response. *Language and cognitive processes*, 28(4):426–442, 2013.

Stefanie E Kuchinsky, Jayne B Ahlstrom, Kenneth I Vaden Jr, Stephanie L Cute, Larry E Humes, Judy R Dubno, and Mark A Eckert. Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*, 50(1):23–34, 2013.

Stefanie E Kuchinsky, Judy R Dubno, and Mark A Eckert. Advances in quantifying listening effort: Growth curve analyses of pupillometry data. *The Journal of the Acoustical Society of America*, 139(4):2101–2101, 2016.

Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016.

Zhen-Hua Ling, Li Deng, and Dong Yu. Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE transactions on audio, speech, and language processing*, 21(10):2129–2139, 2013.

Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen M Meng, and Li Deng. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine*, 32 (3):35–52, 2015.

Paul A Luce. Comprehension of fluent synthetic speech produced by rule. *The Journal of the Acoustical Society of America*, 71(S1):S96–S96, 1982.

Paul A Luce, Timothy C Feustel, and David B Pisoni. Capacity demands in short-term memory for synthetic and. natural speech. *Human factors*, 25(1):17–32, 1983.

Carol L Mackersie and Natalie Calderon-Moultrie. Autonomic nervous system reactivity during speech repetition tasks: Heart rate variability and skin conductance. *Ear and Hearing*, 37:118S–125S, 2016.

LM Manous and DB Pisoni. Effects of signal duration on the perception of natural and synthetic speech. *Research on Speech Perception Progress Report No*, 10, 1984.

Ronan McGarrigle, Kevin J Munro, Piers Dawes, Andrew J Stewart, David R Moore, Johanna G Barry, and Sygal 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(7):433–440, 2014.

Thomas Merritt, Robert AJ Clark, Zhizheng Wu, Junichi Yamagishi, and Simon King. Deep neural network-guided unit selection synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference*, pages 5145–5149. IEEE, 2016.

Daniel Mirman. *Growth curve analysis and visualization using R*. Chapman and Hall/CRC, 2017.

Masanori Morise, Fumiya Yokomori, and Kenji 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.

Tobias Neher, Giso Grimm, and Volker Hohmann. Perceptual consequences of different signal changes due to binaural noise reduction: do hearing loss and working memory capacity play a role? *Ear and hearing*, 35(5):e213–e227, 2014.

Howard C Nusbaum and David B Pisoni. Constraints on the perception of synthetic speech generated by rule. *Behavior Research Methods, Instruments, & Computers*, 17(2):235–242, 1985.

Carol R Paris, Richard D Gilson, Margaret H Thomas, and N Clayton Silver. Effect of synthetic voice intelligibility on speech comprehension. *Human Factors*, 37(2):335–340, 1995.

Carol R Paris, Margaret H Thomas, Richard D Gilson, and J Peter Kincaid. Linguistic cues and memory for synthetic and natural speech. *Human Factors*, 42(3):421–431, 2000.

Jonathan E Peelle. Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, 39(2):204, 2018.

Jonathan E Peelle, Rowena J Eason, Sebastian Schmitter, Christian Schwarzbauer, and Matthew H Davis. Evaluating an acoustically quiet epi sequence for use in fmri studies of speech and auditory processing. *Neuroimage*, 52(4):1410–1419, 2010.

M Kathleen Pichora-Fuller, Sophia E Kramer, Mark A Eckert, Brent Edwards, Benjamin WY Hornsby, Larry E Humes, Ulrike Lemke, Thomas Lunner, Mohan Matthen, Carol L Mackersie, et al. Hearing impairment and cognitive energy: The framework for understanding effortful listening (fuel). *Ear and Hearing*, 37:5S–27S, 2016.

Erin M Picou and Todd A Ricketts. The effect of changing the secondary task in dual-task paradigms for measuring listening effort. *Ear and Hearing*, 35(6):611–622, 2014.

Erin M Picou, Todd A Ricketts, and Benjamin WY Hornsby. Visual cues and listening effort: Individual variability. *Journal of Speech, Language, and Hearing Research*, 54(5):1416–1430, 2011.

Erin M Picou, Julia Gordon, and Todd A Ricketts. The effects of noise and reverberation on listening effort for adults with normal hearing. *Ear and hearing*, 37(1):1, 2016.

Tepring Piquado, Derek Isaacowitz, and Arthur Wingfield. Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, 47(3):560–569, 2010.

David B Pisoni, Howard C Nusbaum, and Beth G Greene. Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, 73(11):1665–1676, 1985.

David B Pisoni, Laura M Manous, and Michael J Dedina. Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility. *Computer speech & language*, 2(3-4):303–320, 1987.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE, 2019.

James V Ralston, David B Pisoni, Scott E Lively, Beth G Greene, and John W Mullennix. Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Human factors*, 33(4):471–491, 1991.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

EH Rothauser. Ieee recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, 17:225–246, 1969.

Anastasios Sarampalis, Sridhar Kalluri, Brent Edwards, and Ervin Hafter. Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research*, 52(5):1230–1240, 2009.

Eschman A. Schneider, W. and A. Zuccolotto. *E-Prime User's Guide*. Psychology Software Tools, Inc., Pittsburgh, 2012.

Maximilian Schwalm, Andreas Keinath, and Hubert D Zimmer. Pupillometry as a method for measuring mental workload within a simulated driving task. *Human Factors for assistance and automation*, (1986):1–13, 2008.

Diemo Schwarz. A system for data-driven concatenative sound synthesis. 2000.

Scott Seeman and Rebecca Sims. Comparison of psychophysiological and dual-task measures of listening effort. *Journal of Speech, Language, and Hearing Research*, 58(6):1781–1792, 2015.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj 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)*, pages 4779–4783. IEEE, 2018.

Olympia Simantiraki, Martin Cooke, and Simon King. Impact of different speech types on listening effort. *Proc. Interspeech 2018*, pages 2267–2271, 2018.

Gerit P Sonntag, Thomas Portele, and Felicitas Haas. Comparing the comprehensibility of different synthetic voices in a dual task experiment. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

John Sweller. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier, 2011.

Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4784–4788. IEEE, 2018.

Paul Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.

K Tokuda, T Yoshimura, T Masuko, T Kobayashi, and T Kitamura. Duration modeling in hmm-based speech synthesis system. In *Proc. of ICSLP*, volume 2, pages 29–32, 1998.

Keiichi Tokuda and Alan W Black. The Blizzard Challenge-2005: Evaluating corpus-based speech synthesis on common datasets. In *Interspeech*, pages 77–80, 2005.

Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. Speech parameter generation from hmm using dynamic features. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 660–663. IEEE, 1995.

Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki, and Takao Kobayashi. Hidden markov models based on multi-space probability distribution for pitch pattern modeling. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, pages 229–232. IEEE, 1999.

Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1315–1318. IEEE, 2000.

Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *SSW*, 125, 2016.

Savvas Varsamopoulos, Koen Bertels, and Carmen Almudever. Designing neural network based decoders for surface codes. 11 2018.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. Grammar as a foreign language. In *Advances in neural information processing systems*, pages 2773–2781, 2015.

Anita E Wagner, Paolo Toffanin, and Deniz Başkent. The timing and effort of lexical access in natural and degraded speech. *Frontiers in Psychology*, 7:398, 2016.

Yang Wang, Adriana A Zekveld, Dorothea Wendt, Thomas Lunner, Graham Naylor, and Sophia E Kramer. Pupil light reflex evoked by light-emitting diode and computer screen: Methodology and association with need for recovery in daily life. *PloS one*, 13(6), 2018.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

Oliver Watts. Ophelia. `git@github.com:CSTR-Edinburgh/ophelia.git`, 2019.

Oliver Watts, Gustav Eje Henter, Jason Fong, and Cassia Valentini-Botinhao. Where do the improvements come from in sequence-to-sequence neural tts? In *10th ISCA Speech Synthesis Workshop. ISCA, Vienna, Austria (September 2019)*, 2019.

Dorothea Wendt, Torsten Dau, and Jens Hjortkjær. Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in psychology*, 7: 345, 2016.

Mirjam Wester, Martin Corley, and Rasmus Dall. The temporal delay hypothesis: natural, vocoded and synthetic speech. *Proceedings DiSS Edinburgh, UK*, 2015.

Conor J Wild, Afiqah Yusuf, Daryl E Wilson, Jonathan E Peelle, Matthew H Davis, and Ingrid S Johnsrude. Effortful listening: the processing of degraded speech depends critically on attention. *Journal of Neuroscience*, 32(40):14010–14021, 2012.

Matthew B Winn, Dorothea Wendt, Thomas Koelewijn, and Stefanie 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*, 22:2331216518800869, 2018.

Stephen J Winters and David B Pisoni. Perception and comprehension of synthetic speech. *Research on spoken language processing report*, 26:95–138, 2004.

Patricia Wright and Daniel Kahneman. Evidence for alternative strategies of sentence retention. *The Quarterly journal of experimental psychology*, 23(2):197–213, 1971.

Yu-Hsiang Wu, Nazan Aksan, Matthew Rizzo, Elizabeth Stangl, Xuyang Zhang, and Ruth Bentler. Measuring listening effort: Driving simulator vs. simple dual-task paradigm. *Ear and hearing*, 35 (6):623, 2014.

Zhizheng Wu, Oliver Watts, and Simon King. Merlin: An open source neural network speech synthesis system. *Proc. SSW, Sunnyvale, USA*, 2016.

Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 492–498. IEEE, 2021.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Sixth European Conference on Speech Communication and Technology*, 1999.

Dong Yu and Li Deng. Deep learning and its applications to signal and information processing [exploratory dsp]. *IEEE Signal Processing Magazine*, 28(1):145–154, 2010.

Kai Yu and Steve Young. Continuous f0 modeling for hmm based statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1071–1079, 2010.

Adriana A Zekveld and Sophia E Kramer. Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*, 51(3):277–284, 2014.

Adriana A Zekveld, Sophia E Kramer, and Joost M Festen. Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and hearing*, 31(4):480–490, 2010.

Adriana A Zekveld, Sophia E Kramer, and Joost 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.

Adriana A Zekveld, Dirk J Heslenfeld, Ingrid S Johnsrude, Niek J Versfeld, and Sophia E Kramer. The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *Neuroimage*, 101:76–86, 2014.

Adriana A Zekveld, Thomas Koelewijn, and Sophia E Kramer. The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends in hearing*, 22:2331216518777174, 2018.

Heiga Zen and Andrew Senior. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 3844–3848. IEEE, 2014.

Heiga Zen, Keiichiro Oura, Takashi Nose, Junichi Yamagishi, Shinji Sako, Tomoki Toda, Takashi Masuko, Alan W Black, and Keiichi Tokuda. Recent development of the HMM-based speech synthesis system (HTS). In *APSIPA*, pages 121–130, 2009.

Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, pages 7962–7966. IEEE, 2013.