Statistical Parametric Speech Synthesis Using Conversational Data and Phenomena

Rasmus Dall

Doctor of Philosophy

Institute for Language, Cognition and Computation

School of Informatics

University of Edinburgh

2017

Abstract

Statistical parametric text-to-speech synthesis currently relies on predefined and highly controlled prompts read in a “neutral” voice. This thesis presents work on utilising recordings of free conversation for the purpose of filled pause synthesis and as an inspiration for improved general modelling of speech for text-to-speech synthesis purposes. A corpus of both standard prompts and free conversation is presented and the potential usefulness of conversational speech as the basis for text-to-speech voices is validated. Additionally, through psycholinguistic experimentation it is shown that filled pauses can have potential subconscious benefits to the listener but that current text-to-speech voices cannot replicate these effects. A method for pronunciation variant forced alignment is presented in order to obtain a more accurate automatic speech segmentation something which is particularly bad for spontaneously produced speech. This pronunciation variant alignment is utilised not only to create a more accurate underlying acoustic model, but also as the driving force behind creating more natural pronunciation prediction at synthesis time. While this improves both the standard and spontaneous voices the naturalness of spontaneous speech based voices still lags behind the quality of voices based on standard read prompts. Thus, the synthesis of filled pauses is investigated in relation to specific phonetic modelling of filled pauses and through techniques for the mixing of standard prompts with spontaneous utterances in order to retain the higher quality of standard speech based voices while still utilising the spontaneous speech for filled pause modelling. A method for predicting where to insert filled pauses in the speech stream is also developed and presented, relying on an analysis of human filled pause usage and a mix of language modelling methods. The method achieves an insertion accuracy in close agreement with human usage. The various approaches are evaluated and their improvements documented throughout the thesis, however, at the end the resulting filled pause quality is assessed through a repetition of the psycholinguistic experiments and an evaluation of the compilation of all developed methods.

Acknowledgements

Thanks to my supervisors Simon King, Junichi Yamagishi and Mirjam Wester. Without whom this thesis would not have happened. Thanks to my collaborator on the NST project - Marcus Tomalin - whose support and insights greatly helped this work. To Martin Corley for insisting I get the stats right and for help setting up the psycholinguistic studies. To Sandrine Brogneaux for a fruitful collaboration on pronunciation variation. To Keiichi Tokuda for steadfast support, the willingness to review my work on a yearly basis and for being a gracious host during my visit at his lab in Nagoya. To all the people at the CSTR - of which there are too many to name - don't stop the karaoke! To all the people at Nitech SP Lab for a great time in Japan. To the people at Google for letting me try a stupid idea while they paid me for it. To my examiners, Jürgen Trouvain and Hiroshi Shimodaira, for taking the time to engage with my work in a positive manner. To all the fantastic colleagues around the world for engaging discussions whenever we met, conferences, visits or the like. Thanks again to Mirjam for ensuring I was always on track and for engaging with my work above and beyond the call of duty. Thanks to all the people who's supported me during the work and put up with my inability to engage with other things because I was thinking about my work. In particular, thanks to my family back in Denmark who barely saw me during this time. Thanks to Hamish Kallin for distracting me with Zelda and whisky, to Johnny Horn and Jonathan Riise for distracting me with beer and whisky, to Zac Laporte for distracting me with whisky buying and drinking, to Fraser Cormack for distracting me with board games and beer, to Tom Merrit for fun times making radio.

Thanks to all the other people who've supported me throughout the time I've been doing this work that I've failed to mention.

This work was supported by the JSTCREST uDialogue project.

And whisky. And beer. And people.

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.

(*Rasmus Dall*)

Table of Contents

List of Figures

List of Abbreviations

AMT	Amazon Mechanical Turk
ASR	Automatic Speech Recognition
BAP	Bandwise Aperiodicity
CPU	Central Processing Unit
CSS	Conversational Speech Synthesis
DM	Discourse Marker
DNN	Deep Neural Network
DT	Decision Tree
DP	Decision Parameter
EM	Expectation-Maximisation
F0	Fundamental Frequency
f-RNN	Full Output Layer RNN
FP	Filled Pause
FSA	Finite State Acceptor
FST	Finite State Transducer
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HSMM	Hidden Semi-Markov Model
HTK	Hidden Markov Model Toolkit
HTS	H Triple S HMM Speech Synthesis System
Hz	Hertz
IPA	International Phonetic Alphabet
kHz	Kilohertz
LM	Language Model
logF0	Logarithmic Fundamental Frequency
MDL	Minimum Description Length
MCEP	Mel Cepstral Coefficient

MOS	Mean Opinion Score
NLP	Natural Language Processing
NN	Neural Network
OOV	Out-of-vocabulary
PCFG	Probabilistic Context Free Grammar
POS	Part-of-speech
RP	Received Pronunciation
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SLF	HTK Standard Lattice Format
SP	Silent Pause
SPSS	Statistical Parametric Speech Synthesis
SR	Speaking Rate
SVM	Support Vector Machine
TTS	Text-to-Speech Synthesis
WER	Word Error Rate

List of Data Sets

AMI Corpus A corpus of meeting recordings orthographically transcribed. See Carletta (2007) for details. Used to provide sentences for the reaction time experiments in Chapter 4 and 10. Also used for the corpus for predicting filled pause insertion in Chapter 9 and the text-based task in Section 3.3.

Arctic Corpus A corpus of phonetically balanced sentences from out of copyright novels for the purpose of recording TTS databases. See Kominek and Black (2003) for details. Used for the Read Speech Corpus (see Chapter 3) and for the text-based task in Section 3.3.

British Telephone Speech Corpus An unreleased corpus of recorded telephone conversations faithfully transcribed. Used for the corpus for predicting filled pause insertions in Chapter 9.

Change Detection Sentences A set of sentences developed for use in change detection experiments kindly provided by Martin Corley and used in Section 4.6.

CereProc Sentences A number of natural read and spontaneous sentences recorded for TTS kindly provided by CereProc for use in the first test of Section 3.3.

Combilex Dictionary A pronunciation dictionary for TTS containing pronunciation variants utilised in Chapter 6 for word reductions and variation and in Chapter 5 for pronunciation variant alignment.

Fisher Corpus A corpus of recorded telephone conversations faithfully transcribed. See Cieri et al. (2004) for details. Used for the corpus for predicting filled pause insertions in Chapter 9.

Read Speech Corpus A corpus of read speech from a Female British English speaker recorded for the purpose of this thesis. See Chapter 3 for details. This corpus is in parallel to, and from the same speaker, the spontaneous speech corpus.

Spontaneous Speech Corpus A corpus of spontenaoues speech from a Female British English speaker recorded for the purpose of this thesis. See Chapter 3 for details. This corpus is in parallel to, and from the same speaker, the read speech corpus.

Parallel Read and Spontaneous Sentences 50 read, and 50 spontaneous sentences of the same content recorded for the purposes of this thesis. See Chapter 3 for details. Used throughout the thesis for synthesis evaluation, as the basis for Gold standard in Chapter 5 and also in the acoustic only study in Section 3.3.

Switchboard A corpus of recorded telephone conversations faithfully transcribed. See (Goodfrey et al., 1992) for details. Used for the corpus for predicting filled pause insertions in Chapter 9.

WSJ Corpus A corpus of news articles from the Wall Street Journal. See (Paul and Baker, 1992) for details. Used for the human filled pause insertion investigation in Section 9.3 and the text based task in Section 3.3.

Chapter 1

Introduction and Thesis Structure

Classically, all Text-To-Speech synthesis (TTS) systems aim to deliver clear, intelligible and natural speech in a neutral read speaking style. The current state-of-the-art TTS systems are considered to produce speech as intelligible as naturally produced human speech and, although the naturalness lags behind, it is still quite reasonable (King and Karaiskos (2009); Clark et al. (2007); Fraser and King (2007); Karaiskos et al. (2008)).

Such neutral read speech is appropriate for many applications, in particular those in which TTS is used for information systems such as GPS car navigation, but in recent years interest has grown in a number of applications that require synthetic voices with a conversational style. These applications include new voices for speech-impaired people, interactive life-like conversational agents, speech-to-speech translation, personal assistants, robots and indeed most places in which human-computer interaction is done through speech. These applications require more than simply being understood, we want these applications to interact with, and potentially act like, humans in as natural a way as possible; and when humans communicate via language it is primarily through conversational speech and not by reading aloud from a book. The most natural form of language is arguably speech, not text, and so our systems should sound like they are *speaking* not *reading*.

This leads to the central hypothesis of this thesis:

> **Main Hypothesis:** The use of spontaneous conversational data and phenomena can lead to improvements in the perception of TTS output.

This is a very broad claim, however, and therefore this thesis will investigate a number of claims derived from the Main Hypothesis – all of which, if supported, may provide support for the Main Hypothesis. In particular, this thesis will investigate the following hypothesis':

> **Hypothesis 1:** Spontaneously produced natural conversational speech is considered more natural than read aloud prompts.
>
> **Hypothesis 1a:** TTS voices can benefit from the use of spontaneous speech in the training corpus.
>
> **Hypothesis 1b:** TTS voices can benefit from being more spontaneous - whether based on read or spontaneous speech.
>
> **Hypothesis 2:** Pronunciation variation is important both to model and to realise conversational speech.
>
> **Hypothesis 2a:** Standard read speech-based voices can also benefit from pronunciation variation.
>
> **Hypothesis 3:** Filled pauses can provide benefits to the listener in TTS.
>
> **Hypothesis 3a:** Filled pauses can be usefully employed in a TTS system if properly realised.
>
> **Hypothesis 3b:** Filled pause insertion can be accurately predicted from text.

This thesis is structured around investigating these claims. The rest of this introductory chapter will give a brief overview of the contents of this thesis with a particular focus on how each of the chapters relates to the hypotheses and the individual consequences of each claim.

Note how several of the hypotheses explicitly mention improvements in read speech-based voices.[1] This is important as it is a guiding element in some of the methods proposed that they should also potentially benefit standard read speech-based voices – in addition to the spontaneous speech-based ones. The primary example of this is the linguistic feature sets based on parsing proposed in Chapter 8. The linguistic feature sets are beneficial for read speech-based voices, but not so much for spontaneous speech-based voices because of the difficulties encountered in parsing transcriptions of spontaneous speech. Nevertheless, the inclusion of parsing-based feature sets in this

[1] Note that what a read speech-based voice is will be properly defined in Chapter 2. Briefly, it refers to the standard type of TTS voice created based on read aloud prompts.

thesis is fitting as the resulting improvement to sentence level prosody has a decidedly conversational component associated with it, in addition to the more obvious benefit of filled pause modelling.

Chapter 2 provides an overview of the literature on the use of conversational data and phenomena in TTS and a brief introduction to the TTS methods used in this thesis. It will provide the necessary background to understand why the use of conversational data and phenomena is of interest, but also why it constitutes a challenge for conventional TTS systems. I will start by defining what is meant by “conversational” or “spontaneous” speech and how this is not simply a matter of speaking style, but of speaking mode. Next, I’ll describe the type of conversational phenomena that will be focused on – filled pauses – and why these are of particular interest. This is followed by the aforementioned review of prior literature. Finally, the chapter will give a brief introduction to the TTS, particularly Hidden Markov Model (HMM) based speech synthesis, which is the main focus of this thesis, and describe how these systems currently produce speech and where they may be deficient when dealing with the type of data used in this thesis.

Chapter 3 will present a corpus of speech designed to allow for the testing of the main hypothesis, and provide motivation for the paths chosen in this thesis by presenting a preliminary study into the naturalness of spontaneous conversational data. A standard evaluation method of TTS is Mean Opinion Score (MOS) naturalness tests. A study is presented using this method to compare natural read and spontaneous speech, highlighting the potential naturalness gains from using spontaneous data over conventional read speech data – directly testing Hypothesis 1. Following that a study directly comparing a voice based on either read or spontaneous speech is presented, in which it is shown that such a spontaneous speech-based voice using filled pauses and discourse markers is dis-preferred compared to one based on read speech. Showing that Hypothesis 1a is not immediately supported.

While naturalness is a standard metric in TTS, it is not the only possible metric of progress. In psycholinguistics, filled pauses (FPs) have been much investigated and several potential benefits have been uncovered. These benefits, amongst other studies, provide motivation for the focus on FPs in this thesis. Firstly an overview of the acoustic effects of FPs is presented, showing effects such as lowered F0 and longer durations, this is followed by some psycholinguistic investigations. These psycholinguistic approaches rely on different measurements than standard TTS, and the potential benefit of using FPs in TTS is investigated using a series of psycholinguistic experiments based around the unconscious benefit to the listener of their use – testing Hypothesis 3. These studies investigate peoples' reaction times and change detection rates, and compare the effect of FPs in natural human speech to vocoded and synthetic speech. It is demonstrated that there are potential unconscious benefits of FPs in natural speech, but that neither vocoded nor synthetic speech based on standard read corpora fully exhibit the same benefits.

Together, Chapter 3 and 4 provide evidence that current techniques and corpora cannot convincingly synthesise disfluencies nor properly model conversational speech. The synthesis of disfluencies from read speech suffers from a lack of training samples (none exist in standard corpora) and the overall quality of a conversational speech-based voice is lower than a voice based on read speech. One important reason for this is to be found in the increased variability of the spontaneous data, making it difficult for standard methods to model it effectively. Forced alignment is one such method which has difficulties dealing with spontaneous speech. In Chapter 5, this is illustrated by analysing the output of the standard forced alignment method which shows that many additional, and much more serious, errors occur when aligning spontaneous speech compared to standard read speech, particularly with regard to reductions and deletions. A method for improving this alignment is presented and evaluated, testing Hypothesis 2, showing that fixing some of these errors in a consistent manner is beneficial. While this is indeed useful, it is also shown that the quality of both read and spontaneous speech should improve if a method for producing such reduced phonetisations is found. An improvement to read speech-based voices through pronunciation reduction would also support Hypothesis 2a as reductions and deletions occur more frequently in spontaneous speech and should thus give the output a more conversational character. Chapter 6 presents a method for producing variant and reduced pronunciations for synthesis. By utilising word probability information from language modelling and phone sequence probabilities from the training data, gradeable reduced output synthesis more consistent with the training data can be produced. Although simply deterministically producing reduced pronunciations showed the most promise. In this way, consistency between training and synthesis can be maintained while also producing a better underlying model.

We now have a method for automatic pronunciation variant alignment and synthesis of reduced speech. While applying this method increases the naturalness of the output speech, it does not, however, directly improve the synthesis of conversational phenomena as needed to support Hypothesis 3; and unfortunately voices based on spontaneous speech, while improved, still do not match the quality of voices based on read speech. Thus Chapter 7 presents and discusses how FPs may be better realised by utilising specific phone modelling and data mixing techniques – testing Hypothesis 3a. The standard dictionary entry for FPs provides a few phonetic choices, but due to the frequency of occurrence and special nature of FPs it is quite possible that a separate phone model is appropriate. Therefore, a specific separate phone for FPs is introduced. Additionally, a number of data-mixing techniques are presented which aim to retain the quality of read speech-based voices while benefiting from the better FP modelling obtainable by using spontaneous speech. The data-mixing techniques also improve the modelling of FP-specific phenomena such as fundamental frequency (F0) lowering and phone lengthening.

Conversational phenomena occur in circumstances outside of normal sentence structure. For this reason, in Chapter 8, it is hypothesised that parsing may help identify them as being outside of this normal structure and that this can further help the realisation of FPs. Furthermore, parsing should have generally beneficial effects on TTS output by improving sentence level prosody – likely pertinent to the feeling of conversationality as well. Using Probabilistic Context Free Grammar (PCFG) and Dependency Structure parsing, an above-word level phrase structure can be obtained which, as well as providing useful new linguistic features for both read and spontaneous speech, identifies FPs as existing in a special phrase structure. It is shown that this further improves the output speech quality and FP synthesis.

We now have methods for improved disfluency synthesis, but knowing how to synthesise disfluencies is only useful if we know when to synthesise them. It is assumed in the previous chapters of this thesis, that the input text contains FPs, however, normally written text does not contain disfluencies and so a method for predicting their usage is presented. This directly tests Hypothesis 3b. In Chapter 9, such a method for automatically inserting FPs and discourse markers (DMs) into the output speech is presented, together with an evaluation of the resulting predictions compared to the performance of humans on the same task. This method allows for control of the degree of disfluency via a disfluency parameter controlling the number of inserted FPs and DMs, thus allowing the output to be as disfluent, or fluent, as necessary.

As a final evaluation, all the methods developed in this thesis are compared to the original voice type in Chapter 10, where I will also revisit the psycholinguistic experiments of Chapter 4, testing if the methods presented in this thesis have an unconscious, besides conscious, effect. The chapter ends with a presentation of a tool released as part of this thesis which can perform most of the modifications to the TTS pipeline proposed in this thesis. Finally, Chapter 11 presents an overall discussion of the contributions of the thesis, potential future work and a few final remarks.

1.1 Scope of Thesis

There are a very high number of potential issues with conversational speech, as highlighted in Chapter 2, and an equally large number of conversational phenomena which could be explored. It has therefore been necessary to limit the scope of this thesis severely. Filled pauses has been chosen as the representative conversational phenomena due to the large body of psycholinguistic literature and its potential to provide measures of TTS quality and improvements of a more subtle nature than the currently used metrics. In fact the particular type of FP that is the focus of the thesis, mid-sentential FPs is a further limitation of scope as FPs can serve a few different purposes in a conversation. In Chapter 4 this and other phenomena are discussed and pointers for the interested reader is provided as to where to begin investigating those, however, they are not dealt with in this thesis. The focus on pronunciation variation was chosen over other issues for a few reasons. Pronunciation variation is one of the more obvious difference between read and spontaneous speech, it is also a very representative example of the increased variability of spontaneous speech as compared to read speech. Through better modelling of pronunciation variation at both training and synthesis time a large part of the full TTS pipeline is affected, and it is a necessary first step to allow for the modelling of many other phenomena such as prosodic variation, as without proper modelling of individual phones in the right context, hoping to get intonation, emphasis and other prosodic variables right seem very difficult. Further investigation of the exclusive use of conversational data for TTS after Chapters 5 and 6 was not performed, although it could have, as this would have forced the thesis away from the discussion and investigation of filled pause use, it is hoped that future researchers will further this element (see Chapter 11 for idea for future directions. One way of evaluating the use of filled pauses, and conversational data, in TTS which has not been presented is the use of a real-world interaction with a life-like agent. Such work was considered out-of-scope of this thesis as it is necessary to establish the use of conversational data and pheneomena in their own right before evaluating them in such a scenario - particularly due to the difficulties encountered when modelling such speech. Such evaluations, while out of scope of this thesis, was considered by the present author and initial investigations into this was carried out by (Dreke, 2016) - an MsC thesis supervised by the present author. The interested reader should refer to this.

1.2 A Note About Collaborative Work.

No one is an island, neither am I. As is clear from the publication list below, I have had many collaborators while doing the work in this thesis. Therefore, throughout this thesis, each chapter will be prefaced with a note, similar to this, which will detail who, besides the main author, has also contributed to that chapter and what they have contributed, besides also detailing the contributions of the main author.

1.3 Papers Published as Part of This Work

During the course of the creation of this thesis, a number of peer-reviewed papers have been published. Some exist in rewritten form as chapters or parts of chapters in this thesis, while some have not been included. Below is a list of all papers published during work on this thesis, and in each chapter it is detailed which, if any, papers the contents of that chapter is based on/expanding upon.

First-Authored Papers:

- Dall, R., Yamagishi, J., and King, S. (2014). Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation. In *Proc. Speech Prosody*, Dublin, Ireland.

- Dall, R., Wester, M., and Corley, M. (2014). The Effect of Filled Pauses and Speaking Rate on Speech Comprehension in Natural, Vocoded and Synthetic Speech. In *Proc. Interspeech*, Singapore.

- Dall, R., Tomalin, M., Wester, M., Byrne, W., and King, S. (2014). Investigating Automatic & Human Filled Pause Insertion for Speech Synthesis. In *Proc. Interspeech*, Singapore.

- Dall, R., Wester, M. and Corley, M. (2015). Disfluencies in Change Detection in Natural, Vocoded and Synthetic Speech. In *Proc. Disfluencies in Spontaneous Speech*, Edinburgh, Scotland, UK.

- Dall, R., Brognaux, S., Richmond, K., Valentini-Botinhao, C., Henter, G. E., Hirschberg, J., Yamagishi, J., and King, S. (2016). Testing the Consistency Assumption: Pronunciation Variant Forced Alignment in Read and Spontaneous Speech Synthesis. In *Proc. ICASSP*, Shanghai, China.

- Dall, R., and Gonzalvo, X. (2016). JNDSLAM: A SLAM extension for Speech Synthesis. In *Proc. Speech Prosody*, Boston, USA.

- Dall, R., Hashimoto, K., Oura, K., Nankaku, Y. and Tokuda, K. (2016). Redefining the Linguistic Context Feature Set for HMM and DNN TTS Through Position and Parsing. In *Proc. Interspeech*, San Francisco, USA.

- Dall, R., Tomalin, M. and Wester, M. (2016). Synthesising Filled Pauses: Representation and Datamixing. In *Proc. SSW 9*, Sunnyvale, USA.

Co-Authored Papers:

- d'Alessandro, N., Tilmanne, J., Astrinaki, M., Hueber, T., Dall, R., Ravet, T., Moinet, A., Cakmak, H., Babacan, H., Barbulescu, A., Parfait, V., Huguenin, V., Kalayc, S., and Hu, Q. (2013). Reactive Statistical Mapping: Towards the Sketching of Performative Control with Data. In *Innovative and Creative Developments in Multimodal Interaction Systems*, Rybarczyk, Y., Cardoso, T., Rosas, J., and Camarinha-Matos, L. M. (eds.), Springer, New York.

- Aylett, M., Dall, R., Ghoshal, A., Henter, G. E., and Merritt, T. (2014). A Flexible Front-End for HTS. In *Proc. Interspeech*, Singapore.

- Tomalin, M., Wester, M., Dall, R., Byrne, B., and King, S. (2015). A Lattice-Based Approach to Automatic Filled Pause Insertion. In *Proc. Disfluencies in Spontaneous Speech*, Edinburgh, Scotland, UK.

- Wester, M., Corley, M., and Dall, R. (2015). The Temporal Delay Hypothesis: Natural, Vocoded and Synthetic Speech. In *Proc. Disfluencies in Spontaneous Speech*, Edinburgh, Scotland, UK.

- Wester, M., Aylett, M., Tomalin, M., and Dall, R. (2015). Artificial Personality and Disfluency. In *Proc. Interspeech*, Dresden, Germany.

Chapter 2

Background

2.1 Preliminary Definitions: Speech, Modes and Styles

Before properly beginning it is important to frame the discussions in this thesis with regards to some of the terms used.

Often, when speaking about language, we divide it into the spoken and written form. Spoken language is anything that usually comes under the heading of “speech”. That is, it has been uttered aloud by some speaker of some language, whereas written language defines anything which has been written down. This divide is evident within language research, as there are almost separate fields of study for the spoken and the written. Natural Language Processing (NLP) deals with text, written language, and Speech Technology (or Processing), as the name suggests, with spoken language – despite many overlapping interests/techniques/etc.

For the purposes of this work, this is not generally satisfactory and a different demarcation of language is needed. Text can be read aloud and speech can be written down; and both can still retain defining features. It is therefore useful to demarcate speech into two *mediums* - written and spoken - and two *modes* - planned and spontaneous (more on modes below). Note that this is not meant to represent any ultimate truths about language but rather to make useful distinctions for the purposes of this thesis. This gives us a four-way distinction between written text, transcribed speech, read text and spontaneous speech (see Figure 2.1). Of course one could argue that there is a continuum of various speech modes between read text and spontaneous speech such as a newsreader’s speech, lecture speech and theatre manuscripts - this is certainly true but not relevant for current purposes. What is relevant here is that the important distinction is not between the written and spoken medium, but rather between the planned and

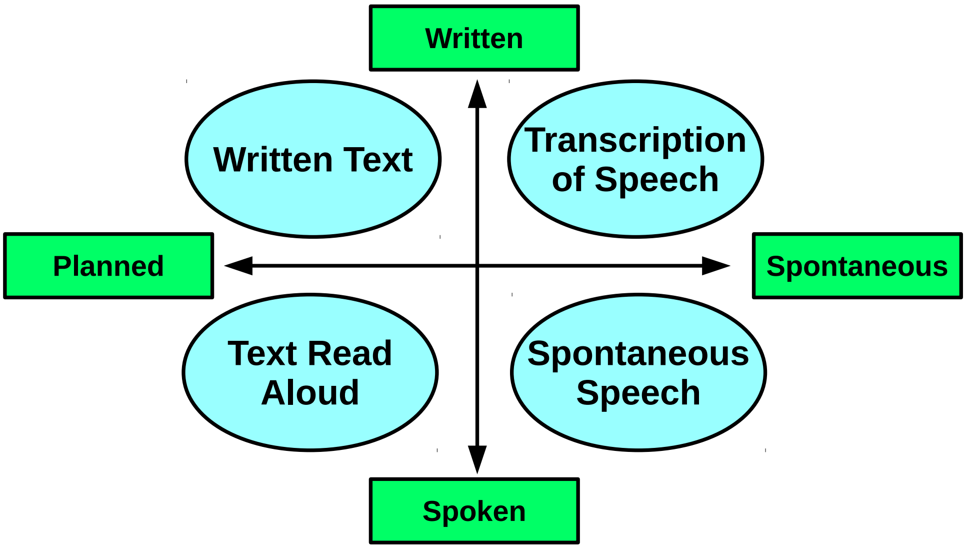


Figure 2.1: Divisions of language in the work. Note that this is not an absolute division and that a continuum exist between the shown divisions such as lecture speech, newsreader's speech and theatre manuscripts.

spontaneous modes. Therefore what I will often simply call "read" speech may refer to both the written text and the read-aloud versions of the same, and conversely what will be called "conversational" or "spontaneous" speech may refer to both spontaneously uttered sentences in a (relatively) unconstrained setting and its faithfully transcribed text.

Secondly, we need to distinguish between modes and styles of speech. With this I mean to say that there is a hierarchical difference between a mode and a style. In the speech synthesis literature there is an interest in the synthesis of various speech styles, with a particular focus on synthesising emotions (e.g., Yamagishi et al. (2005); Tachibana et al. (2005); Schröder (2001)). However, once the focus turns to conversational speech synthesis this is also often seen as a style of speech (as in Adell et al. (2010a, 2012)) or considered a matter of getting the emotional, or phatic, utterances right (as in Campbell (2006a)) – that is, getting the style right. This is perhaps not intentional by the authors of work on conversational and emotional speech, however it manages to confuse a simple truth. Conversational speech is not a style of speech such as emotions are, rather conversational speech is a *mode* of speech within which various speaking *styles* can be expressed. Read speech can be angry or sad, and so can conversational speech. The mode encompasses the style, and within a mode one can utilise styles. Therefore in this work the term speaking styles will be used to refer to speech wrought in a certain style such as the angry, sad, sarcastic or happy style[1] and speaking modes will be used to refer to the more general ways of speaking within which we can employ speaking styles.

2.2 Conversational Speech and Phenomena in Earlier TTS Literature

While conversational speech synthesis (CSS) seems desirable, it has received only sporadic direct attention. Most research has been done within the concatenative or diphone synthesis paradigm (e.g., Adell et al. (2007a); Sundaram and Narayanan (2002); Campbell (2007); Andersson et al. (2010a)) and to my knowledge only Andersson (2013) and Koriyama et al. (2010) have used Statistical Parametric Speech Synthesis (SPSS) methods in a conversational setting. As the literature is quite sparse, I will here review and present most of the research carried out and make it apparent how this impacts on the current work.

2.2.1 Conversational Speech vs Read Speech

The study of spontaneous speech, as opposed to written text, is not a new field of research, in particular in the linguistics and psycholinguistics literature, and this is because these are different *modes* of language. Differences are present in both spoken and written versions of either mode. If focusing on the spoken mode, some of the differences include that in spontaneously produced speech the speaking rate increases (both at the word (Andersson et al., 2010b) and syllable (Blaauw, 1992) level) and differs according to the probability of the word (Jurafsky et al., 2001; Bell et al., 2003). Read speech has steeper pitch movements (Blaauw, 1992), and in conversational speech more reduced pronunciations are observed (Nakamura et al., 2006; Bell et al., 2003), similarly there's a spectral space reduction for vowels (van Son and Pols, 1999) and also more variance is observed in the pronunciation of all phones in spontaneous speech (Nakamura et al., 2008).

[1] It may seem more appropriate for some of these to call them "tones of voice" or "emotional speech" instead of style, and as such style of speech and tone of voice/emotionaly speech could also have been used. However to be consistent with existing literature style is used for emotions etc. and mode for conversational/read.

Other differences can also be in the written form as well as the spoken, one of the most noticeable being the introduction of disfluencies such as repeats, restarts, repairs and filled pauses (as observed in e.g. Fox Tree (2000); Clark and Wasow (1998); Lickley (1996)). If we discard such disfluencies, we find other structural differences. Andersson et al. (2010b) notes that, even in speech data cleaned for use in synthesis, many simple differences are observable in approximately the same amount of data – in conversational speech there are less word types, differing words are more frequent and also Nakamura et al. (2006) notes a differing part-of-speech (POS) class distribution. Syntactically the two also differ. Notably, conversational speech contains no punctuation (read speech implicitly contains the punctuation from the text), this must be inferred from the speech signal or the structure of the speech in the form of e.g., sentence-like units (Liu et al., 2006). Furthermore, differing grammatical constructions are used, often with the use of discourse markers (also called lexical fillers) which differ in written and spontaneous text (Heeman and Allen, 1999), and often “wrappers” with little propositional content are used (Campbell, 2006b). In sum the differences are many and varied and, perhaps unsurprisingly, most present challenges to current speech synthesis methods.

Current HMM-systems rely on read aloud pre-written prompts which are recorded under very controlled conditions in sound studios. The various speech parameters necessary for synthesis (more details in Section 2.3.2) are extracted and HMM models are trained on these. Unfortunately, such recordings are very bad examples of how an actual spontaneous conversation sounds, these recordings are exactly the kind of read speech which is so different from spontaneous speech, and so the trained models are *learning the wrong mode of speech* (see Chapter 3 for further discussion of this). From the observed differences above it is clear that, e.g., durational models will be directly mislearned, conversational speech is faster and so the read speech durations learned will be mismatched. Chapter 4 highlights one potential problem with this – namely slower sentence recognition by listeners. The reduction of the spectral space and increased pronunciation variation is likely to affect both the spectrum but also the fundamental frequency (F0) modelling of the speech. Furthermore, we can observe that due to the difference in word use and the inclusion of disfluencies, the decision tree-based context clustering will be affected as well. Contexts (e.g., phonetic, part-of-speech (POS) or word) which are common in read speech may be different in conversational speech and so the decision tree will group contexts differingly and new contexts are likely to become salient. Not only in the training of the models do we have challenges however. At synthesis time other issues appear. Text analysis elements rely on cues not present in conversational speech, such as in phrase-break prediction where punctuation is the most reliable predictor used, and language models used for POS-tagging are also trained on read speech and do not deal well with, e.g., disfluencies (Bulyko et al., 2007). These are some of the problems that must be overcome in order to successfully synthesise conversational speech, and unfortunately these are too many problems to deal with in one thesis and only a subset will be attempted to be dealt with here.

2.2.2 Hidden Markov Model-based Conversational Synthesis

In Andersson et al. (2010b, 2012); Andersson (2013), the synthesis of speech with conversational characteristics was done in a data-driven way using Hidden Markov Model (HMM) synthesis. Two primary methods were investigated. For the first method, the core HMM-system was not modified and an HMM voice was trained on carefully recorded and selected sentences which were spontaneously produced. To elicit the spontaneous utterances a voice talent was recorded in a studio, first standard read speech prompts were recorded, followed by a free conversation with the experimenter - a corpus collection method similar to that used in this thesis (see Chapter 3). This resulted in an amount of data sufficient to build a voice purely on the conversational data. For the second method, a blending technique developed for style blending (Yamagishi et al., 2005) was used to mix the modes, in which an additional context - read or conversational - was added to the data annotations, and both conversational and read data were pooled when building the voice. The blended voice was made in order to increase the phonetic coverage of the resulting voice as the conversational data contains less unique quinphone types than the read, and the read data, conversely, was prosodically less rich than the conversational data.

Andersson et al. (2010b) found that the conversational voice synthesised sentences containing “fillers” – that is discourse markers, repetitions and filled pauses – was better than a voice based on read speech in terms of both naturalness and “conversational style”, and furthermore, that the blended voice lay somewhere in between. While this shows the importance of using the right data for HMM synthesis, it is not clear that the voice blending technique actually improved the resulting synthesis or simply blended the benefits – see Chapter 7 for more on this.

There are two further notable results, firstly, as Andersson et al. (2012) notes, there is a correlation between the conversationality ratings and the naturalness ratings of listeners. This suggests that we must be careful when using naturalness as an evaluation metric and I will discuss this further in Section 3.3. Secondly, when synthesising utterances containing “fillers” the spontaneous voice was rated more natural than the read voice, however when these were removed there was a preference shift towards favouring the read aloud voice. The simple removal of a few words had a clear effect. In fact, the “clean” sentences still contained many words and expressions only commonly found in conversational speech thus this tendency suggests that it is not only voice prosody and characteristics which determine what people consider conversational but rather that word choice also has a significant impact, this will be discussed further in Chapter 3.3. It is also worth noting that compared to the approach by Adell et al. (Adell et al., 2006, 2012, 2010b) a disfluency is not directly modelled but rather treated as a word in itself and consequently duration modelling and prosody is derived in the standard HMM approach.

Another approach to HMM-based CSS was taken by Koriyama et al. (2010) who used average-voice HMM-synthesis (Yamagishi, 2006) to overcome the data sparsity problem. A comparison of average-voice models built on conversational or read speech and then adapted with conversational data was made, and here it was found that a large read voice adapted to a small amount of conversational data was better than a small conversational average voice adapted with conversational data. Through a two stage adaptation technique, an equivalent naturalness to the large read voice was achieved using the conversational data, however this method relied on a large amount of conversational data, which is expensive to obtain. In Koriyama et al. (2011) the effect of new linguistic features for decision tree clustering on a conversational voice was investigated, and it was shown that additional contexts such as phone prolongation, disfluency and tone label improved the resulting voice’s naturalness, however the resulting voice was not compared to a standard read voice. One noticeable thing about the conversational data is that there is more prosodic variation in the speech. To enable a higher variation in prosody prediction, Koriyama et al. (2012) created a prosodic-unit HMM-based F0 prediction module, in which the X-JTobi (Maekawa et al., 2002) labelling scheme was used as the base unit. The method successfully synthesised increased F0 variation and decreased the size of the decision tree to 40% of the conventional method, though perceptual tests did not show a significant preference for the method. The use of additional features will also be investigated in this thesis, in Chapter 8, however those will be a markedly different set of features than those used by Koriyama et. al.

2.2.3 Insertion of Conversational Phenomena

In early attempts at inserting conversational phenomena such as laughter, breathing and filled pauses, Sundaram and Narayanan (2002) showed, using a limited domain synthesiser, that the insertion of these conversational phenomena - called VoiceFonts after Campbell (1998) - made the resulting synthesised sentences more than three times as likely (from 8% to 27%) to be confused for natural speech in a four-way forced choice test. Based on this, Sundaram and Narayanan (2003) present a method for text transformation inspired by a psycholinguistic analysis and implemented partly by a statistical method through Finite State Acceptors (FSA) and heuristic rules. A newer paper, Cadic and Segalen (2008), inserts laughter and hesitations using unit selection in a similar way by simply adding data that has the desired features. While such methods seem relatively simplistic it is notable that text analysis methods applied in many TTS systems today do not even attempt to predict breathing and filled pauses but merely phrase breaks, which may or may not include breathing, depending on context. This despite earlier efforts showing the feasibility of synthesising breathing and other phenomena by expanding the dataset used. A series of experiments by Adell, Bonafonte and Escudero (Adell et al., 2006, 2007a, 2010a, 2007b, 2012) show the feasibility of synthesising disfluent speech using unit selection synthesis. After analysing repetitions and filled pauses in over ten hours of speech, Adell et al. (2006) adopts an approach with duration and pitch contour prediction, which is further extended in Adell et al. (2007b), based on modifying the overall prosodic predictions for the equivalent fluent sentence and locally changing the disfluent sentence prosody. That is, the disfluency is analysed using the definition by Shriberg (1994) and Levelt (1983). This definition splits the disfluent utterance into three elements, the reparandum, editing phase and repair. The reparandum is what comes *before* the disfluency, the editing phase is the disfluency itself and the repair what follows the disfluency. Adell et al. (2012) illustrates this well (p. 54 – see also Figure 4.1 in Chapter 4 of this thesis) and their method works by an underlying fluency model in which two sentences' prosody is synthesised (the utterance prior and post disfluency), and then the prosody is modified according to the disfluency and local context. Adell et al. (2007a) predicts the insertions of filled pauses and Adell et al. (2010a) demonstrates an equivalent Mean Opinion Score (MOS) for the disfluent system compared to a standard system, showing that it is possible to insert filled pauses into synthetic speech without degrading its quality. The method was also evaluated in a pragmatic scenario, represented by a simulated card game, in which the disfluent system was far preferred over the non-disfluent system, thereby demonstrating the ecological validity of the insertion of disfluencies. In Chapter 9, further discussion of disfluency insertion prediction is presented together with another method for predicting disfluencies.

2.2.3.1 "Canned" Insertion

All the above approaches have in many ways relied on an extension of earlier data and this approach is taken to the limit in Campbell (2006a,b, 2007) where conversational characteristics are inserted at the phrase-level in a unit selection system. In the analysis of the ESP corpus, a large corpora of everyday speech consisting of over 1500 hours of recordings (Campbell, 2004), a distinction between A-type and I-type content in an utterance was made (Campbell, 2006b), with the former for the purpose of expressing affect and the latter for information.[2] The suggestion was then to synthesise the I-type content via normal unit selection and the A-type content through a phrase-based method, i.e., by selecting longer "units" corresponding to a full A-type utterance. In this framework A-type utterances act as "wrappers" around the I-type "fillers" [3] and they are also the "wrappers" which Andersson et al. (2012) (above) remove from the "real" sentence. No evaluation of the resulting system in Campbell (2007) was done as:

> [E]ach utterance synthesised is a complete and self-contained natural-speech segment. There is no concatenation, except at the phrase level, where utterances are separated naturally by pauses, no prosody modification, and no signal processing. By definition each utterance is natural. (p.26)

While this is in some sense true it goes against the goals of speech synthesis, and is a step toward what is called "canned speech". Which is essentially not *synthesis* but rather *playback* of recordings. The classic problem for canned speech synthesis also applies to this phrase-level approach. We simply cannot record every possible utterance of any language. It is therefore telling that – despite the enormous size of the ESP corpus (1500 hours) – A-type utterances account for more than half the corpus (Campbell, 2007) and despite the repetitive and more limited nature of A-type utterances, the corpus still does not cover every possible scenario. Consider the possible textual variances of an A-type utterance like "this might seem obvious" Campbell (2007), we can

[2] Campbell himself admits this to be an over-simplification.

[3] Note the atypical use of "fillers" to denote the words carrying the actual propositional content.

quickly name a few - "this should be obvious", "obviously", "this might seem superfluous", "this could seem obvious" - then consider that each of these can be realised prosodically in many ways and one quickly realises the infeasibility of the approach.

2.2.4 Duration and Pronunciation Selection for CSS

In a series of studies, Werner, Eichner and colleagues highlights the importance of durational modelling (Eichner et al., 2002; Werner et al., 2004) and pronunciation selection (Werner et al., 2004; Werner and Hoffmann, 2006, 2007; Eichner et al., 2003). These are important for CSS as the increased speaking rate and higher pronunciation reduction rate of spontaneous speech must be modelled somehow. Jurafsky et al. (2001) notes that the language model (unigram, bigram and reverse bigram) probability of a word affects both its duration (probable words are shorter) and its reduction rate (probable words are more often reduced) and this has been extended in Bell et al. (2003). This was picked up by Eichner et al. (2002) who implemented a multigram language model-based durational modelling in a diphone-based system, and found it improved the perceived quality of the synthetic speech. It was also proposed that durational modelling and pronunciation variation selection go hand in hand; in Eichner et al. (2003) a pronunciation variant lexicon was utilised together with the LM probabilities in order to control word reductions to speed up the speech. Werner et al. (2004) unified the two approaches. These are summarised and further listening tests and comparisons to other system performances and to unmodified natural speech were presented in Werner and Hoffmann (2006, 2007), in which the proposed method achieves both a higher mean opinion score (MOS) and lower listening effort than the standard method (though not as high as natural). While all this work was done with the, now outdated, diphone synthesis method, these are effects which are only indirectly modelled in current HMM systems. The standard approach to phone reduction is to simply assume it will occur if a reduction is frequent and stable enough that the context dependent HMMs will pick up on the reduction and model it naturally. In Chapters 5 and 6, this assumption is tested using pronunciation variant forced alignment and pronunciation variant synthesis, respecitvely.

2.2.5 Related Research Areas Review

While research in Conversational Speech Synthesis (CSS) is directly relevant, three other main areas of research deal with similar issues to CSS. These are Emotional/Expressive

Speech Synthesis (ESS), Automatic Speech Recognition (ASR) and NLP. Emotional speech synthesis deals with the synthesis of various emotional states of the speaker and the work here is also relevant to CSS as the goals are closely tied, namely to make speech synthesis sound more natural by applying effects observed in natural spontaneous speech. ASR deals with conversational data all the time both as training data and as input data. In NLP, conversational speech is dealt with at the written level; however it still provides many challenges and areas such as POS-tagging, Semantic Role Labelling, Syntax Trees and Machine Translation have all begun looking at actual conversational data.

2.2.5.1 Emotional, Expressive Speech Synthesis

Schröder (2001) reviews, now older, attempts at synthesising emotions and in Schröder (2009) this review is updated with newer approaches. All methods of synthesis have been used for ESS, that is both formant synthesis (Burkhardt and Sendlmeier, 2000), diphone synthesis (Montero et al., 1999), unit selection (Eide et al., 2004) and HMM-synthesis (Miyanaga et al., 2004; Yamagishi et al., 2005). The main developments in ESS are similar to the conversational systems with most current methods relying on aquiring the right data. It is notable however that with the move from unit selection to HMM-synthesis it is no longer purely the data that determines what can be done, but that various techniques such as style control (Miyanaga et al., 2004), mixed modelling (Yamagishi et al., 2003, 2005) and model adaptation (Yamagishi et al., 2007) allow for increased control over the output due to the parametric representation. These methods generally rely on a smaller amount of data from each style and modify a few of the parameters of the speech in a data-driven manner. Average-voice speaker adaptation (Yamagishi et al., 2007) works by adapting some parts of the HMM-models (e.g. F0, spectrum and duration) to a target speaker using a transformation function. The use of such a method for CSS is also investigated in Chapter 7. Another direction for ESS is personality modelling. Trouvain et al. (2006) manipulates the perceived personality of the synthetic speech through directly setting parameters, such as F0, speaking rate and loudness, in a diphone synthesis system. Showing that prosodic modelling can influence the listeners' perception of synthetic speech. In a study related to the work presented in this thesis, Wester et al. (2015a) show that filled pauses can also influence peoples' perception of the speakers' personality. Finally we got the synthesis of affective sounds such as laughter (Cakmak, 2016; Trouvain and Schröder, 2004; Campbell, 2006a), backchannels Oertel et al. (2016) and similar (Eide et al., 2004), which all relate to the current thesis (see Chapter 4 for some more detail).

2.2.5.2 ASR and NLP

In ASR, the focus has long been on conversational data, as this is the base mode of input to the system for both training and recognition, I will therefore only touch upon research which has an impact on CSS. In ASR, conversational phenomena provide a challenge and a choice, do we attempt to recognise all these phenomena or only what is important for the content of the utterance? “Reconstruction” of the intended sentence by removal of conversational characteristics (Fitzgerald and Jelinek, 2008; Fitzgerald, 2009) is an example of the latter, however the main focus has been on identification of these characteristics in the ASR output (Nakatani and Hirschberg, 1994; Liu et al., 2006; Stolcke et al., 1998) or letting the language model handle them (Huang and Renals, 2010; Heeman and Allen, 1999; Bulyko et al., 2007). Removal, identification and modelling of these issues are also a focus in NLP where they are employed for various other purposes. When POS-tagging conversational speech, it is questionable if current tagsets are adequate. Charniak and Johnson (2001) present a method for detecting and removing edits (repetitions, restarts, etc.) as a preparatory step for parsing and similarly Kahn et al. (2004) detect sentence boundaries and interruption points for their parser. Generally, sentence boundary detection is important when dealing with conversational speech due to the lack of punctuation and it has therefore also been investigated in ASR (Gotoh and Renals, 2000). The identification of conversational phenomena is of interest because it provides the option of *generation* of these phenomena. Many identification, or detection, methods can be utilised to train a model which can be used to predict where and when the model phenomena occurs, and as such we might be able to *insert* conversational markers into speech in the read style. Using statistical machine translation (SMT), methods Neubig et al. (2012) proposed a way to create transcripts cleaned of conversational phenomena. They called it speaking style transformation and it was implemented using a weighted finite state transducer network. As the name suggests, the method can potentially be used not only to create clean transcripts from the original, but the method can also be turned on its head to produce transcripts containing the conversational phenomena. One problem with such an approach is that SMT can produce unforeseen results and as such the output may need further processing before it is useful (Shriberg, 1994). One way of ensuring that such insertions are done in a meaningful manner is to utilise language models (LMs) to “smooth” the output in an “over-generate and re-rank” style, a method best known in natural language generation, where it has been employed to ensure more fluent output from generation systems (Belz, 2005), and machine translation, where LMs are used to rank potential translations (Zhang et al., 2006). In the current context, LMs capable of handling conversational speech could be useful. A further way of performing such a text transformation would be to utilise paraphrasing techniques (Callison-Burch, 2008; Barzilay and McKeown, 2001; Ganitkevitch et al., 2011; Madnani and Dorr, 2010; Woodsend and Lapata, 2011). A paraphrase is an alternative way of expressing one phrase by means of another. Paraphrasing techniques thus attempt to automatically identify appropriate paraphrases of any given phrase. Three types of paraphrases are generally considered - lexical synonyms, phrase-level paraphrases and sentential paraphrases (Madnani and Dorr, 2010). All three are potentially interesting by allowing the replacement of words, phrases or entire sentences consisting of generally read speech with their conversational speech counterparts. While such an approach is certainly interesting, it is outwith the scope of this thesis. However, the method for filled pause and discourse marker insertion presented in Chapter 9 shares many of the same basic principles and the above discussed idea would be a natural, though laboursome, extension which the results, when investigating the naturalness of read and spontaneous text, in Chapter 3 would suggest to be fruitful.

2.3 A Short Introduction to the TTS Methods Used

The focus of this thesis – with regards to synthesis techniques – is on the statistical parametric speech synthesis (SPSS) technique of hidden markov model-based (HMM) TTS. While unit selection is by no means an unviable approach, the increased flexibility of SPSS allows for operations not possible in a unit selection paradigm, some of which will be employed in Chapter 7. Although many of the experiments and investigations in this thesis could usefully be done using unit selection in addition to SPSS, including unit selection would have widened the scope beyond what was considered reasonable within the constraints of the paradigms and this thesis. I will therefore not present how unit selection works here, and only mention the method in passing where relevant in the thesis. Furthermore, the introduction to SPSS methods here is just that, an introduction. For a more thorough presentation of both unit selection and SPSS methods I recommend Taylor (2009) or a more general textbook on computational linguistic methods like Jurafsky and Martin (2008), and for detailed presentation of the main HMM system used in this thesis, namely HTS, I recommend the technical papers published by the HTS working group at *http://hts.sp.nitech.ac.jp/?Publications* – a good starting point being Zen et al. (2007a).

2.3.1 A Brief History of TTS

Text-to-speech synthesis (TTS) has advanced tremendously in the past twenty years. In the 90s, formant synthesis was replaced with sample-based methods, notably diphone synthesis and unit selection, which lifted the quality of TTS to a standard usable in a multitude of commercial situations, particularly as the intelligibility reached natural speech levels.

In the 00s, statistical parametric speech synthesis (SPSS) methods appeared, most notably HMM-based synthesis (Tokuda et al., 2002), which reduced the footprint of systems to sizes possible to fit on mobile devices (Gutkin et al., 2010) and increased the flexibility of the systems by allowing modifications of the speech, independent of the training corpora (e.g., Tachibana et al. (2005); Yamagishi et al. (2005)).

While SPSS was initially, and surprisingly, judged more natural than unit selection systems in the first Blizzard challenge (Bennett, 2005; Zen and Toda, 2005; Black and Tokuda, 2005) it was soon found that, at sufficiently large dataset sizes, unit selection was superior and therefore remained the "go to" method commercially (Gutkin et al., 2010), whereas most researchers turned towards SPSS.

In recent years, however, the appearance of Neural Network-based (NN) SPSS (e.g., Zen et al. (2013); Wu et al. (2015)) and better vocoders (Agiomyrgiannakis, 2015; Kawahara et al., 1999b; Morise et al., 2016) has closed the gap to make SPSS equivalent in quality to unit selection (Zen and Sak, 2015; Agiomyrgiannakis, 2015). As a consequence, SPSS techniques are increasingly attractive for all TTS applications, particularly as the best unit selection methods tend to be driven by an SPSS model for unit selection (as can be seen in King and Karaiskos (2013)).

Still, SPSS, and TTS in general, is clearly distinguishable from natural speech. There are many remaining reasons for this. Some reasons are that SPSS models are still inaccurate and do not match vocoded speech, vocoded speech is close to, but not quite like, natural speech and the produced prosody is considered flat and unnatural. Each of these reasons has received significant attention in recent years – Merritt et al. (2014) and Henter et al. (2014) provide a summary of how each of these elements limit the final synthesis output.

There is, however, a fourth way in which TTS is unnatural: TTS is modelling and learning to produce neutral read speech. That is, speech produced by a voice talent in a sound-proofed booth reading aloud sets of phonetically balanced sentences. This is, however, not how speech is normally produced. Most speech, truly natural speech, is produced in spontaneous conversations (see Chapter 3 for direct evidence of this), and utilising this type of speech in TTS should be beneficial.

2.3.2 Hidden Markov Model-based Speech Synthesis

HMM synthesis is the primary method of SPSS used in this thesis. It is named such due to the core machine learning technique being employed is the hidden markov model. While NN-based speech synthesis has very recently become the go to method for SPSS research, and is also often employed commercially, this thesis work was started before NN had truly eclipsed HMM synthesis and – as this thesis is not focused on machine learning – it has not attempted to adopt the new method.[4] I will therefore here provide only an overview of HMM, and not NN, synthesis while still referring to it as SPSS synthesis in general.

SPSS has three primary stages: corpus preparation, model training and speech synthesis, illustrated in Figure 2.2. Besides these, two important elements appear throughout: namely, linguistic analysis and signal processing. I will describe these while explaining each of the three stages in creating synthesis output.

In the corpus preparation stage, the corpus is prepared for statistical modelling, in the training stage the statistical model is learned and in the synthesis stage, speech is generated from the learned model. All three stages will be modified in this thesis, however the main focus is on corpus preparation and speech synthesis – primarily realised through linguistic analysis.

2.3.2.1 Corpus Preparation

Corpus preparation generally consists of three elements (Figure 2.3): script preparation, recording and speech segmentation.

The scripts are usually prepared with phone-in-context coverage in mind, the phone being the unit of speech used in SPSS. However, where the phone in language analysis is language independent and fully described through the international phonetic alpha-

[4] Note that NN synthesis is in fact employed in one of the later Chapters (8), however in tandem with HMM synthesis. See also Chapter 11 for a discussion of how the here employed methods would be impacted by the move to NN-based synthesis and how NNs could open up new avenues of research with regards to CSS.

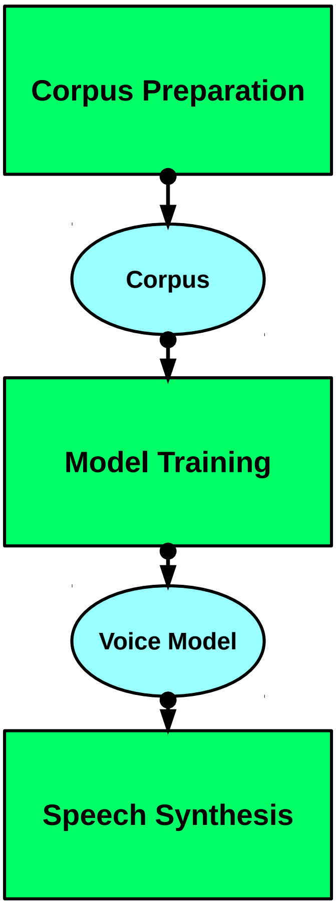


Figure 2.2: An overview of the SPSS workflow.

bet (IPA), the phone in SPSS, is language specific. A phone inventory of a language in SPSS is a description of the minimal pairs of meaningful sounds present in this language. This means that while one can generally map one-to-one from IPA phone to SPSS phone, one SPSS phone may map to several IPA phones if they are not meaningfully different in a given language.[5] A classic example is the phones /l/ and /r/ which are not distinguished between by Japanese speakers, and in a phonetic description of Japanese this would be collapsed to one SPSS phone encompassing both IPA phones. This is indeed the case in the Japanese TTS system OpenJTalk (*http://open-jtalk.sourceforge.net/*) where the phone /r/ is used to represent the two sounds. Another

[5] Note however, that often the term phoneme is used as an alternative. In this thesis phoneme will not be used, although arguably, phone in the SPSS sense is closer to phonemes than to a phone in the IPA sense. However, the literature tends to prefer phone over phoneme so that is the convention adopted in this thesis.

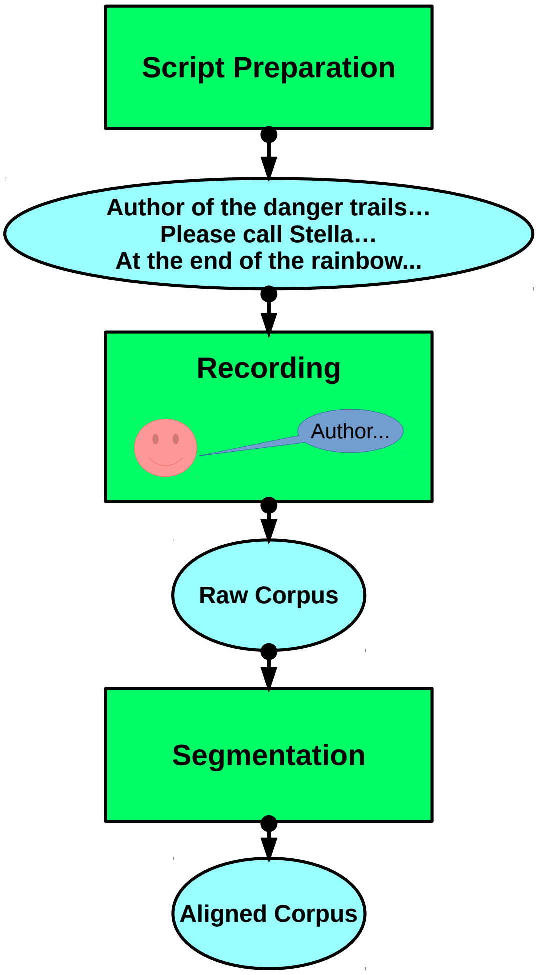


Figure 2.3: An overview of the corpus preparation workflow.

example is the pair of phones /w/ and /v/ in Danish, these do not change the meaning of words but are rather a feature of the speakers' dialect, and would thus be represented by one phone in SPSS. As a consequence of this, a phonetic inventory will also not contain sounds not used in the language in question, for example in English clicks are not used, and are therefore not part of the phonetic inventory of English. Throughout this thesis these sort of phones will be used and not IPA phones and this distinction is important to bear in mind. Specifically, the phone inventory of Combilex (Richmond et al., 2009) for received pronunciation (RP) English will be used.

Once a script is prepared with a sufficient predicted phone coverage (see Kominek and Black (2003) for a detailed discussion of how this can be attained) this script is then recorded. The recordings are usually made in a studio in which each sentence of the script is recorded sequentially at the highest possible quality. Details of this process

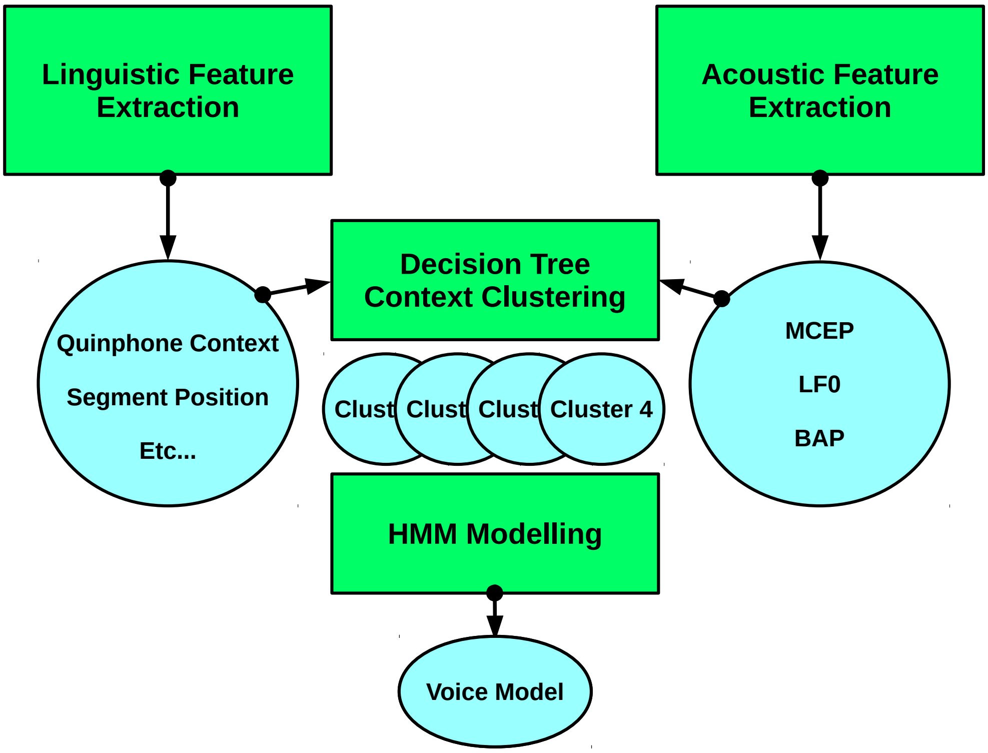


Figure 2.4: An overview of the HMM voice model training workflow.

will be discussed in Chapter 3 where a description of the corpora used in this thesis is given. Next, the recordings are split by utterance, converted to a mono recording and usually downsampled to either 16kHz or, as is increasingly becoming standard, 48kHz, before the speech is segmented into phone units.

This phone segmentation can be done manually, however it is usually done using an automatic method called forced alignment. I will not go into the details of forced alignment here, as Chapter 5 will discuss this at length. It suffices to say here, the method segments each utterance into phones and outputs a time-stamped phone string suitable for use in training.

2.3.2.2 Model Training

The model training stage can be divided into four general elements - linguistic feature extraction, acoustic feature extraction, decision tree context clustering and HMM modelling as illustrated in Figure 2.4.

Linguistic Feature Extraction

Linguistic feature extraction, or textual analysis, is the process of taking input text and producing a marked up string of phones called a full-context label. This is usually done by a system termed the "front-end" as it is often poised as the module applied in the pipeline just prior to the "back-end" (model training) system taking over. This is not entirely accurate. However, for the purpose of this thesis we will keep this nomenclature.

The distinction of "full-context" for the labels refers to the fact that the linguistic description of each phone generally includes the immediate context around it such as the previous and next phone, and position in syllable, word or utterance. In Chapter 8, I will discuss the most common linguistic feature sets used in SPSS at length and in Chapter 10, a research front-end called SiRe, released as part of this thesis, is presented – see those parts of the thesis for details of the workings of a front-end system and its produced analysis.

The astute reader, however, will have noticed that there is an inconsistency in the order of the application of the textual analysis and speech segmentation. In order to do the segmentation a phonetisation of the input text must be produced for the forced alignment process to work, this is indeed usually done by the same system which produces the full-context labels. Depending on the language and the system in question this phonetisation is produced either primarily by dictionary look-up or through grapheme-to-phone (G2P) models. In English, a dictionary is primarily used and G2P rules used as a backup in case of out-of-vocabulary (OOV) words. In this work, however, any OOV word used in either the training or synthesis stage will be added manually to the dictionary to avoid incidental issues with the G2P module which is not what is under examination.

What this means is that the line between corpus preparation and linguistic analysis/model training is more blurred than shown in Figure 2.2. For example, many systems, Festival and SiRe included, will produce a phonetisation to be used for forced alignment which includes markers to allow for later merging with, or re-creation of, the original utterance. In the end, however, what the linguistic feature extraction produces is a full-context label containing a string of phones with a description of that particular phones' context and its duration.

Acoustic Feature Extraction

Acoustic feature extraction consists of taking the input waveform of speech and converting it into a set of acoustic features suitable for modelling. The acoustic feature set usually, and for all systems in this thesis unless otherwise noted, consists of fundamental frequency (F0), mel cepstral coefficients (MCEPs) and bandwise aperiodicity measures (BAP)[6].

MCEPs are used to represent the spectral envelope of the speech, roughly corresponding to the shape of the vocal tract, shape of your lips, position of the tongue etc. The reasons for this are two-fold, partly due to the role of the spectrum in creating the MCEPs and partly due to the source-filter model generally used in SPSS in which the MCEPs represent the filter. MCEPs are made by taking the Discrete Cosine Transform (DCT) of the log of a Mel scaled spectrum (obtained by a Fourier transform of the waveform). The spectrum shows which frequencies, at what intensity/power, are present in the frame (see below) being analysed. This is scaled according to the Mel scale due to the non-linearity of human hearing, in which differences in the lower frequency range are more perceptible compared to the same difference in the higher frequency ranges. The Mel scale makes up for this by making each step equally perceptually relevant. Finally we apply a DCT on the logarithm of the Mel scaled spectrum, this has the effect of decorrelating the coefficients (when doing the Mel scaling the sawtooth bin windows mean there is some overlap) giving us the final MCEP coefficients. Not all of the coefficients are commonly used, specifically the only first 35 coefficients are normally used as they are considered to contain the necessary information (notably though, some of the higher coefficients form the base of several F0 extraction algorithms).

F0 is the fundamental frequency at which the vocal folds vibrate and represent the base tone or pitch of the speaker. Movements in F0 are representative of intonation and is considered the most salient aspect of prosody. Many methods for F0 extraction are available and I will not detail them here. For a comparison of several pitch estimators often used in SPSS (although in a singing synthesis setting) please see Villavicencio et al. (2013). One common thing which is important to note however is that the logarithm of the F0 values from the extraction methods is usually used in the final system as it reduces computational complexity and has the effect of making the F0 distribution more Gaussian (which, as we will see further below, is important), hence I will use the

[6] BAP is the least common feature of these, being generally exclusive to systems using the STRAIGHT (Kawahara et al., 1999b) or WORLD (Morise et al., 2016) vocoders.

term logF0.

BAP is a measure of the aperiodic components of speech binned in frequency bands. As shown in Kawahara et al. (1999a), the modifications generally applied to extract other acoustic features results in the speech signal having a constant F0, a regular harmonic structure and unchanging frication etc., within a small measure of time. However, as speech is not constant and seldom perfectly harmonic, deviations from these introduces inharmonic frequencies – and it is these that are captured by the aperiodicity measure presented in Kawahara et al. (2001). The aperiodic measures are in SPSS, for computational efficiency, then generally binned into 25 bands each representing a single aperiodicity measure.

General for all acoustic features is that they are extracted frame by frame. A frame represents the value(s) of the feature across a window of time. Many window types are possible, but in general the Hamming window is the most common and the one generally applied in speech processing. One can estimate a frame as often and with as wide a window as one likes (bar the sampling rate and sample duration). However, part of the purpose is to reduce the number of variables to estimate and so a frame shift of 5ms with a window of 20-25ms in width is generally used as it is thought that the relevant speech parameters remain relatively stable within that period of time and shorter frame shifts do not make a perceptual difference.

We thus have a frame for every 5 ms of speech with a feature vector consisting of 35 MCEP, 25 BAP and one logF0 measure. This is not quite the final vector as we also use the velocity and acceleration (or delta and delta-delta coefficients) of each measure to improve the standard generation method of the HMM - see the HMM discussion below for why this is necessary.

These acoustic and linguistic features are then used for the decision tree context clustering algorithm.

Decision Tree Context Clustering

Initially proposed by Odell (1995), decision tree-based context clustering is used to cluster "similar" phone models together to provide robust statistics for the HMM to model. Without context clustering, HMM model training would simply consist of an HMM model for each possible full-context label in the training data – generating an unwieldy number of models with no generalisability to unseen contexts.

For decision tree-based context clustering, the simplifying assumption is made that logF0, MCEP and BAP measures are independent of each other, and can thus be mod-

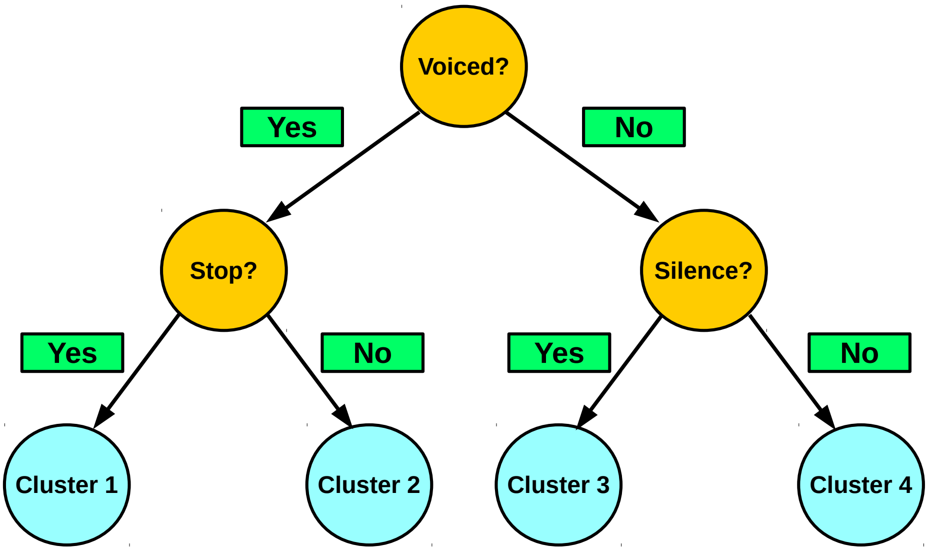


Figure 2.5: A small sample decision tree.

elled independently (Henter et al. (2014) show this does not fully hold). Thus, three separate decision trees are created – one for each measure.

Figure 2.5 illustrates toy a sample decision tree. A set of questions derived from the possible features in the linguistic description of each phone is used to split the data into clusters. At each node, starting at the top, all questions not yet used higher up in the tree are asked about each phone. The phones are clustered into two halves depending on the answer and the “quality” of the split is measured. The algorithm is a greedy algorithm meaning it will pick the best choice locally at each node not considering consequences further down the tree. Thus, the split of the highest “quality” is picked at each node, and the process is then repeated for each new resulting node until some stopping criterion is reached. At this point each node contains a number of context models representing one cluster of linguistically and acoustically similar phones.

As growing the tree fully is not of much use, as it will result in one context in each cluster, a stopping criterion must be employed. Two criteria are normally applied, and if either applies growing the tree further from the node in question is stopped and a leaf is created. The first criterion is a minimum occupancy count, defining the minimum number of data points in a cluster – if a cluster would end up containing less than this amount, growing is stopped. Normally minimally 10 contexts for each cluster is used. The other is based on the predicted “quality” gain from the highest “quality” split. If

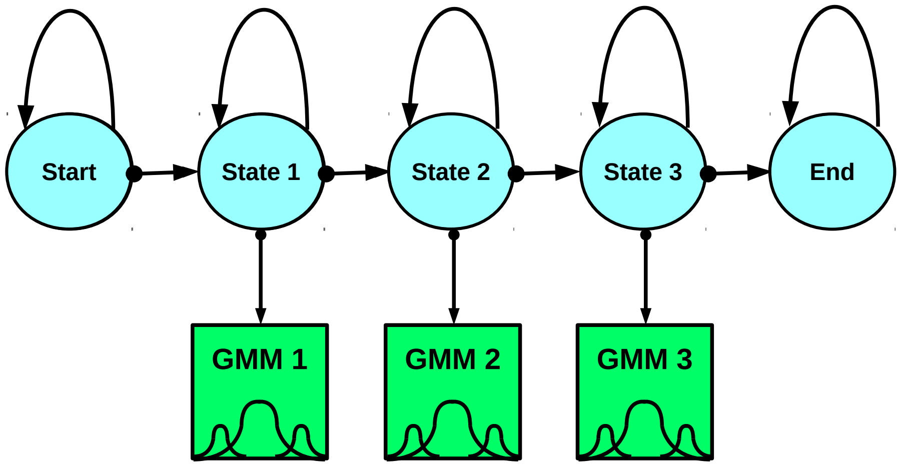


Figure 2.6: A model 3-state HMM.

it is less than some threshold, growing is stopped. "Quality" can be defined by many differing metrics, however, they all aim to measure the benefit of splitting the data in two based on the particular question. Normally in TTS the Minimum Description Length (MDL) criterion is used to determine the quality of a split. Essentially, what the MDL criterion does is to calculate if splitting the data in two enables a shorter code length for describing the data by measuring the number of bits necessary to encode either the data split on this question or not splitting on this question. A shorter code equals a better split.

Besides the three acoustic trees, a duration tree is also created, based on durations for each phone (obtained through alignment) in order to predict phone durations at synthesis time. This is preferred over estimating duration directly from the HMM's transition probabilities – leading to HMM TTS effectively being Hidden *Semi*-Markov Modelling (HSMM, Zen et al. (2007b)) despite normally being called HMM TTS.

2.3.2.3 HSMM Modelling

We now have a number of decision trees clustering all phone contexts into a number of differing clusters depending on which acoustic measure is under consideration. For each cluster, for each acoustic measure, an HMM is then trained.

An HMM model in TTS encompasses the idea of speech as beads on a string. It is a directed acyclic graph consisting of a number of states connected in a forward moving manner (Figure 2.6) such that each state is connected to itself, its previous and next state, but when moving through the states one can only either stay in the same state or transition to the next, one can never go back (i.e. a left-to-right HMM is used in speech). Each state contains a probability distribution (generally a Gaussian Mixture Model – GMM) representing the training data aligned to that state. It is a Markov model because it is assumed that the current observation only depends on the current state, and it is hidden because, in the general case, we can only observe the output generated by the states but do not know the state sequence which generated this output (in TTS we know this when generating from the models, see 2.3.2.4, but not when training the model).

In TTS, we model each full-context phone with an HMM consisting of five emitting states[7] and a mixture of multivariate gaussians in each state – trained on the statistics of the acoustic features extracted from the phones in the given cluster that the HMM represents.

Once we have our set of context clusters each associated with a single HMM we can estimate the state transition probability matrix A and state probability distributions B of each HMM using the Baum-Welch algorithm[8]. The fundamental idea is that we have a sequence of speech frames, our observation sequence:

$$O = (o_1, o_2, ..., o_n) \tag{2.1}$$

where n is the number of frames in the sequence, and we need to find out which sequence of HMM states, Q, generated this observation sequence:

$$Q = (q_1, q_2, ..., q_n) \tag{2.2}$$

which will allow us to estimate the model $M = (A, B)$. A can be calculated by obtaining the maximum likelihood estimate of the probability of the transition between states i and j. this estimate is calculated by counting the number of times (*numtrans*) this particular transition occurs and then normalising it by the total number of times any transition from state i was taken (remember that this will either be to the next state or to stay in the current state):

$$trans_{ij} = \frac{numtrans(i \rightarrow j)}{\sum_{q \in Q} numtrans(i \rightarrow q)} \tag{2.3}$$

B by taking the observation vectors aligned to each state and using those to calculate the probability density function (PDF) of the state – in synthesis, a GMM.

[7] Actually seven states, however, the first and the last are empty dummy states.

[8] The explanation here is shorter but generally similar to that of Jurafsky and Martin (2008) upon which it is based.

Unfortunately, we only know the observation sequence and not the state sequence, this is the hidden part, so we cannot directly calculate these values. The first insight of the Baum-Welch algorithm is that we can iteratively estimate them. We do this by using an initial estimate for the model and then use this estimate to derive progressively better model. The second insight is that we can obtain our estimated probabilities by computing the likelihood of a particular observation and then distributing the probability mass among all the different paths (state sequences) that contributed to this.

We find this by using the forward and backward algorithms. The forward algorithm calculates the probability of being in state i at time t and the backward algorithm the probability of seeing the observations from time $t+1$ onwards assuming we are in state i at time t[9]. We can get the probability of being in state i at time t and transitioning to state j by using the output of these two algorithms (here denoted $fw_t(i)$ and $bw_t(i)$), the state transition a_{ij} and the observation probability $b_j(o_{t+1})$ as given by the equation below:

$$transprob_t(i,j) = \frac{fw_t(i) * a_{ij} * b_j(o_{t+1}) * bw_{t+1}(j)}{fw_T(N)} \tag{2.4}$$

The total expected number of transitions is then the sum over all t of $transprob$. The maximum likelihood probability estimate of transitioning from state i to j can then be found by normalising this value by the total expected number of transitions from state i. Which we get by summing over all transitions out of state i:

$$transest_{ij} = \frac{\sum_{t=1}^{T-1} transprob_t(i,j)}{\sum_{t=1}^{T-1} \sum_{j=1}^{N} transprob_t(i,j)} \tag{2.5}$$

For recalculating the state PDFs we need an estimate of the probability of being in state i and observing a feature vector v. We can obtain this estimate by finding the probability of being in state i at time t in which we observed the feature vector, v, and dividing it by the probability of being in state i at time t in general:

$$observest_i(o) = \frac{\sum_{t=1 s.t. o_t=v}^{T} stateprob_t(i)}{\sum_{t=1}^{T} stateprob_t(i)} \tag{2.6}$$

Where $stateprob_t(i)$ is the probability of being in state i at time t given by the forward and backward probabilities and the total observation probability given the model:

$$stateprob_t(i) = \frac{fw_t(i) * bw_t(i)}{P(O|M)} \tag{2.7}$$

[9] Essentially the sum of the probability of all paths leading to state i at time t from either the front or the back of the observation sequence, see Jurafsky and Martin (2008) Chapter 6 for details.

The Baum-Welch algorithm thus consist of three steps. Initialisation - where the initial model parameters of the model M is set (usually in a flat-start setting). Expectation - computing $stateprob_t(i)$ and $transprob_t(j,i)$ for all t, i and j. Maximisation - recompute the model parameters A and B using $transest_{ji}$ and $observest_i(o)$ for all transitions ji and observation vectors o. The Initialisation step is only run once, whereas the Expectation and Maximisation steps are run iteratively until convergence. This iterative nature makes the Baum-Welch algorithm an EM (Expectation-Maximisation) type algorithm. This yields a final trained HMM model for each phone cluster usable for synthesis.

It is worth noting here that in modern TTS systems we do not use the "vanilla" version of the HMM as presented here. A number of improvements have been found and are generally applied, such as the previously mentioned use of a Hidden Semi-Markov Model (Zen et al., 2007b) and others such as the trajectory model (Zen et al., 2004) – many additional papers on the development of HTS is to be found at the HTS website *http://hts.sp.nitech.ac.jp/?Publications*.

2.3.2.4 Synthesis

Synthesis consists of three steps - linguistic feature extraction, HMM model decoding and waveform reconstruction (see Figure 2.7).

The linguistic feature extraction works exactly as in the training step, yielding a string of full-context phones. For each phone, we can then find the appropriate HMM model by traversing the decision trees and concatenating all HMM models by connecting the start and end states with a transition between the models.

HMM model decoding is then the process of finding the most likely path through the HMM states and generating the appropriate feature vector. As mentioned previously we employ an explicit duration model, and this model is used to determine the overall duration of each phone, thus the most likely path through the HMM states must respect this overall constraint. A specialised version of the Baum-Welch algorithm is used to do this taking into account both the dynamic features produced by the decoding and also the multiple mixtures as presented in Tokuda et al. (2000). This will not be detailed here.

Once each full-context phone model has been decoded for each acoustic stream we are left with a set of frames, each with a vector of acoustic features of MCEPs, BAPs and F0 measures. At this point the method of waveform reconstruction consists of a reversal of the acoustic feature extraction, in order to recombine the differ-

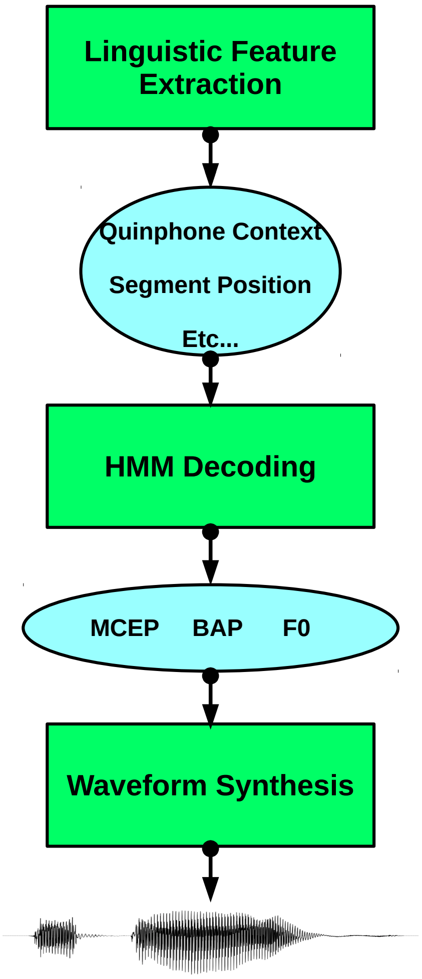


Figure 2.7: The HMM synthesis pipeline.

ent features into a waveform. In this thesis, the process of waveform generation will generally be performed by the STRAIGHT vocoder (which is also used for feature extraction), unless otherwise noted. I will not detail how STRAIGHT works; the interested reader can find more detail and papers by referring to the STRAIGHT website *http://www.wakayama-u.ac.jp/ kawahara/PSSws/* – a good starting point being Kawahara et al. (1999b).

2.4 Chapter Conclusions

In this chapter the background necessary to follow and understand the choices and work made in this thesis has been presented. Initially a discussion of speaking styles and modes highlighted that this thesis deals with the two modes of read and spontaneous speech and that this can cover both the spoken and written medium. Then, an overview of previous literature dealing with conversational speech and phenomena in TTS has been presented followed by a brief overview of how HMM TTS works with references for further reading. In the next chapter, we will turn our focus toward the creation of suitable corpora for testing the hypotheses of the thesis and an initial investigation into the naturalness of read and spontaneous speech is presented.

Chapter 3

The Naturalness of Read and Spontaneous Speech Recorded for TTS

A Note About Collaborative Work

The corpus of read and spontaneous speech presented in this chapter was prepared and recorded by me, orthographically transcribed by Alexandra Delipata and Zachary Palmer Laporte, and their transcriptions checked for consistency by me. The work presented in Section 3.3 is based on Dall et al. (2014c) authored by me and co-authored by my supervisors who provided valuable insights in the process. For the direct comparison between a read and spontaneous speech-based voice, Mirjam Wester and Marcus Tomalin provided valuable discussion and helped define the corpus from which the synthesised samples were drawn from.

3.1 Introduction

In order to do TTS using conversational data and phenomena it is necessary to *have* conversational data to base it on. As no corpora of conversational data recorded directly for TTS are currently publicly available, I will in this chapter present a methodology for obtaining a suitable corpus. The corpus needs to be parallel in the sense that the speech data must come from the same voice talent and be recorded under the same conditions using the same recording tools and contain recordings of both read and spontaneous speech. Without such parallel data from the same speaker one cannot meaningfully compare between the resulting synthesis based on either read or spontaneous speech data. Such a parallel corpus is here presented and subsequently a MOS analysis of the naturalness of the obtained naturally produced spontaneous and read speech will be presented in order to verify that spontaneous speech data can potentially provide direct naturalness benefits to synthetic speech.

A TTS corpus must, traditionally (see e.g., *http://www.festvox.org/festvox/c2176.html*), fulfil the following criteria:

- It must be of sufficient size.
- The recordings must be at a high enough sampling rate.
- As little background noise as possible must be present.
- It must be from one speaker with no overlapping speech (from e.g. interviewers).
- It should preferably be phonetically balanced.

While a number of corpora of conversational speech exist, amongst others the Switchboard (Goodfrey et al., 1992), Fisher (Cieri et al., 2004) and AMI (Carletta, 2007) corpora, none of these are of the required recording quality, they contain speech from several speakers, often overlapping, and the amount of speech from each speaker is limited. These are issues which cannot be ignored in standard TTS corpora. The corpus, to my knowledge, closest to meeting the necessary criteria is the Buckeye corpus (Pitt et al., 2007) which contains studio recordings of an hour of speech from 40 people having a conversation with the interviewer. Unfortunately, some issues arise when considering these recordings. Firstly, the interviewer can often be heard in the background of the recordings, this results in segments of overlapping speech unsuitable for synthesis, which have to be removed. The recordings also contain too much background noise for useful modelling, and with a sampling rate of 16khz it lies at the lower end of the scale of what would normally be used[1]. While Andersson (2013); Andersson et al. (2012, 2010b,a) did present and use several corpora of conversational speech, these are unfortunately unavailable for public use, the same accounts for the small amount of data borrowed from Cereproc used in the first part of Section 3.3 and therefore a corpus for use in this thesis was created.

[1] While 16khz has been standard for many years, 44 or 48khz are becoming the standard as illustrated by the move to 48khz in HTS 2.3 demo scripts.

3.2 Recording a Spontaneous Speech Corpus

Recording a spontaneous speech corpus presents several challenges, most notably that of eliciting spontaneous conversational speech in a studio setting to allow for high quality recordings. Chapter 3 of Andersson (2013) describes his methodology and three recorded corpora, and the methodology employed here closely follows his. The corpora of Andersson (2013) were recorded in a studio in which the experimenter and voice talents, three in all, could see each other through a glass window, such that eye-contact could be retained and bodily gestures seen. In all cases, the conversations were conducted after the recording of standard pre-scripted prompts and no particular restrictions on topic were imposed. The voice talents were, however, instructed not to mimic other people's voices and avoid extreme caricatures of their own (such as highly sarcastic utterances). These conversations resulted in two shorter corpora of about 20 minutes of usable spontaneous speech and one longer of approximately 75 minutes (about 2100 sentences). The longer was used in the mainstay of work by Andersson.

In terms of the above listed criteria for TTS corpora, Andersson's corpora fulfil the requirements for recording conditions. The recordings were made in a studio under controlled circumstances, using a high sampling rate, with minimal background noise and from one speaker with no overlap. The larger 75 minute corpus fulfils the length requirement, with 1 hour being considered a, just, sufficient amount. It is however not certain that it fulfils the phonetic requirement of coverage. Andersson acknowledges this and presents an investigation of the phonetic diversity of the resulting corpus. Comparing to a corpus of read speech consisting of 106 minutes of speech from the same voice talent (20% larger), the read corpora contained 58867 quinphone types and the spontaneous 37564. This is over 50% more quinphone types in only 20% additional speech, demonstrating that, unsurprisingly, the designed manuscripts contain better coverage, something which would theoretically yield a better synthesis model. Andersson et al. (2010b) propose some data mixing techniques exactly because of this lack of phonetic balance (see Chapter 7 for more on these). This is, however, not necessarily an issue, and Lambert et al. (2007) found that randomly selecting sentences for a corpus, versus selecting for phonetic balance, provided better unit selection synthesis. This is probably due to the tendency of random selection to follow the natural distribution in the language and will, generally, provide more data for the most likely things to synthesise - whereas selection based on phonetic balance will provide data less likely to break at edge cases. This shows that we should probably not be too worried about the phonetic balance in a corpus of spontaneous speech, as spontaneous speech will contain primarily the most common phones, and thus should provide good coverage for most standard sentences.

Therefore a method very similar to that of Andersson was followed with few exceptions. A female native British English speaker with prior experience recording TTS corpora, henceforth known as Lucy, was recruited as the voice talent. The recordings took place in a hemi-anechoic chamber using both a headset (Beyerdynamic DT 770 Pro) and a standing microphone (Sennheiser MKH 800 P48). All recordings were done within the timespan of a week and were recorded at 96kHz 32-bit with each microphone providing a single channel.

A manuscript of standard prompts for reading aloud was prepared, consisting of 1982 sentences, 1132 of which comprised the Arctic database (Kominek and Black, 2003), a phonetically balanced set of prompts from out of copyright material developed specifically for speech synthesis, and the last 850 came from sets designed for the Voice Banking project (Yamagishi et al., 2012) and sourced from newspaper material. When recording the read material, the voice talent was given each prompt on a computer screen in front of her. If a sentence was read incorrectly, or in other ways interrupted through coughing etc., she was asked to re-read the sentence and the incorrect material was discarded. The read material was recorded in 30 minute intervals, interspersed with short breaks and recordings of spontaneous conversation. For the spontaneous conversation, the voice talent was seated in the same studio as for the read recordings, and instead of sentences being projected on the screen there was a webcam connection set up such that the interviewer and the voice talent could see each other while talking. This enabled as natural a conversation as possible, without sacrificing recording quality. While the topic was free, the voice talent was encouraged to not put on voices of other people, nor alter their own, and after the first short recorded conversation was also asked to avoid too animated movements as this was noticeable in the recordings of the standing microphone. In total, pre-segmentation, 124 minutes of conversation was recorded. From spontaneous speech from the first day of recording a set of 50 sentences was identified and orthographically transcribed. These 50 sentences were then mixed into the standard read prompts at the next recording session, with the voice talent unaware of this, and recorded as standard prompts. This gave a parallel set of 50 sentences with the same content but in a read and spontaneous rendition. This set of sentences formed the basis for the last part of the investigation into the naturalness of spontaneous versus read speech in Section 3.3.3.

Thus two types of speech were recorded from the same voice talent: traditional phonetically balanced read speech and unscripted spontaneous speech. From here onwards these two sets and types of speech will be referred to as read and conversational or spontaneous, respectively.

3.2.1 Preparing the Speech for Voice Building

Preparing the spontaneous speech for voice building is a notably more involved and expensive process than preparing read speech for voice building. For a read corpus, each sentence is already in a separate sound file with a corresponding orthographic transcription in a text file. One simply needs to confirm that each contains the expected sentence and then the wave files need to be downsampled from 96kHz, 32-bit, to 48kHz 16-bit and either of the microphone channels extracted into mono sound files. The standard TTS pipeline can then be applied. Due to issues not noticed during recording, 9 sentences had to be removed, leaving 1973 sentences of read speech.

For the spontaneous speech, however, we do not have a transcription of the speech, and we do not have each sentence in individual wave files. In order to obtain a useful set of parallel sound and text files, the spontaneous recordings had to be orthographically transcribed and split into utterances based on this transcription. To do the transcription, two transcribers were hired; each transcribed half the material, and their transcriptions were subsequently manually verified by the present author. The transcribers were given the following instructions:

- A sentence should be created when there is a meaningful semantic chunk. However, the aim should be to create sentences of 3-9 seconds in length. If a sentence is much longer than 15 seconds try to see if it can be split in two. If it is over 20 seconds long please always split it in two. Splits should be done where there is silence and preferably where it makes semantic sense for the sentence.

- Transcribe what is being said, not what is meant, unless it is unclear. (e.g. “he hadn’t” not “he had not”, “please came here” not “please come here” or “give me a bear” not “give me a beer”)

- Better to transcribe too many details than too few.

- Add a ‘?’ to the beginning of the sentence if you are unsure of the transcription. In these cases I will make a final judgement.

- Use ‘*’ to mark an unfinished word, if you know what was intended add the word in parenthesis after. (e.g. ”My na*(name) name is”)
- Add an ‘X’ at the beginning of a sentence if sensitive information related to the voice talent is in the sentence. These items will be removed to ensure anonymity for the voice talent.
- Do NOT mark prolongated words. (e.g. if pronounced “weeeeeell” transcribe “well”)
- Transcribe the following filled pauses and backchannels:
  - ‘uh’: er, eh, uh
  - ‘um’: as uh but with audible m ending
  - ‘oh’: considered different from ah
  - ‘ah’: considered different from oh
  - ‘uhu’: as uh but with audible additional u sound (also covers uhuh)
  - ‘uhum’: as uhu but ending with audible m
  - ‘mm’: mm, mhm, mehem
- Transcribe the following non-speech sounds/affective noises:
  - Laughter: if a word goes into laughter transcribe it as “word[laughter]”, if several words in a row have laughter in them transcribe them as “word1[laughter] word2[laughter]”, stand-alone laughter is simply “[laughter]”
  - Coughs: similar to laughter but use tag “[cough]”. Throat-clearing is a cough
  - Lipsmacks: use “[lip]”
  - Grunt: any other unclassifiable sound is “[grunt]”.

After the two transcribers had transcribed the speech, I manually verified all transcriptions and resolved any remaining issues. This resulted in 1829 utterances, however, many of these were not suitable for standard TTS modelling and contained phenomena outside the scope of this thesis. Specifically, all sentences including laughter, coughs, lipsmacks, grunts and unfinished words, that were not at the beginning or end of a sentence or couldn’t be removed from the beginning or end without cutting into the speech were removed. This left a total of 1074 utterances suitable for synthesis.

	Read Arctic	**Read News**	**Spontaneous**
Utterances	1125	848	1074
Words	9979	5432	12607
Phones	37930	19949	42985
Quinphone Types	27947	12870	25840
Triphone Types	9804	5323	8150
Total Duration (m)	66	47	58
Speech Duration (m)	46	34	49
Silence Duration (m)	20	13	9

Table 3.1: Details of the three corpora.

To get a feel for the corpora, an analysis of some of their basic statistics was performed. Based on the audio files, a rough estimate of the size, in minutes, of each corpus was obtained and, based in part on a textual analysis, some phone statistics were also extracted including triphone and quinphones to get an idea of the overall phone context coverage. Table 3.1 summarizes these, and as the amount of read speech recorded is greater than the spontaneous, I have divided the read corpus into two, the Arctic sentences and those from news sources. As can be seen, the spontaneous corpus contains more word tokens, despite fewer utterances, and also more phones in total. However, this does not translate to a larger number of unique quinphone or triphone types as compared to the arctic subset, a similar observation to that of Andersson (2013). The read Arctic and news subsets are of differing sizes, with the arctic set being closely matched to that of the spontaneous. When recording, an attempt was made to match the total amount of spontaneous speech to the total amount of read speech. However, the fact that almost 750 utterances needed to be discarded from the spontaneous corpus means that the corpus is comparable to the arctic subset in size. Therefore, the Arctic subset of the total read corpus will be used in this thesis. It is important that the size of the spontaneous, in terms of actual speech content, is comparable in size to the Arctic corpus, as the size of the Arctic corpus, although today being thought of as a small corpus, is considered an acceptable size. Therefore, despite being shorter than an hour, the corpora recorded here were considered acceptable is size for use in this thesis. The corpora are released under the Apache 2.0 license for all use at the thesis repository (Dall, 2017).

3.3 Naturalness Ratings of Spontaneous and Read Speech

One assumption made in several conversational speech synthesis studies is that spontaneous conversational speech is more natural than read speech (Adell et al., 2010a; Andersson et al., 2012; Campbell, 2007). Thus, it is assumed, synthesis based on conversational speech will similarly increase the system's naturalness. But, it has not been shown that people actually find conversational speech more natural than read speech. The natural intuition certainly is that this is the case. However, earlier studies using spontaneous recordings as the basis for TTS have not managed to increase the perceived naturalness of synthetic speech (Adell et al., 2010a; Koriyama et al., 2010) or at best match it (Adell et al., 2012; Andersson, 2013). It may be that this is due to naturally produced spontaneous speech being considered less natural to listeners, or it may also be due to a difference in the overall quality, not naturalness, of the synthetic speech, which could affect listener ratings.

People can distinguish the two modes of speech with high accuracy despite lexical equivalence (Blaauw, 1991), so it is likely that people will be able to pick up on and judge according to this distinction when asked. One way of testing this is to obtain naturalness ratings of natural speech produced spontaneously in a conversation and when reading aloud from the same speaker. The hypothesis is then, as it has been before, that conversational speech should be considered more natural than read speech.

In speech synthesis research there are two generally used metrics for evaluation, namely intelligibility and naturalness. Intelligibility is a metric for which robust methodologies, such as semantically unpredictable sentences (SUS) (Benoit et al., 1996), have been developed and synthesis systems perform well compared to naturally produced speech (Karaiskos et al., 2008; Clark et al., 2007). While systems tend to perform well on intelligibility they are generally lagging behind naturally produced speech in terms of naturalness.

Naturalness is also a less well-defined concept, although it is generally always used (e.g. in the Blizzard Challenges (King and Karaiskos, 2009; Karaiskos et al., 2008; Fraser and King, 2007)). Naturalness is often considered a measurement of the quality of synthetic speech. However, it is perhaps more precise to describe naturalness as a measure of the *perceived* "naturalness" of synthetic speech due to the way it is measured.

"Naturalness" is in quotes because it is quite possible that as a concept it is underspecified. That is, we do not have an exact definition of what naturalness is. In fact differing studies give participants differing instructions. Naturalness is normally evaluated as a Mean Opinion Score (MOS) where participants rate the quality of the synthetic speech on a 5-point scale ranging from 1-Very Unnatural to 5-Very Natural. However, the Blizzard 2013 evaluation (King and Karaiskos, 2013) instructs participants to give a score which "should reflect your opinion of how natural or unnatural the sentence sounded. You should not judge the grammar or content of the sentence, just how it sounds." In contrast, Adell et al. (2012) explains the meaning of naturalness by asking participants if it is "likely that a person would have said it this way?" (p.470). The two stand in contrast to each other, the one asking to disregard grammar and content, and the other to judge the "way" it was said - easily construed as including content and grammar. If listeners do find it to be underspecified then people's perceptions should be be influenced by their expectations of what naturalness means in any given context. Here, I therefore attempt to influence the prior expectations of listeners by slight variations in instructions to bias them toward either conversational or read speech, and compare this to the general case with no further instructions.

Note that there are genuine worries about the ecological validity of MOS-scale naturalness tests of isolated sentences presented under isolated listening conditions. It is not the purpose of this investigation, nor this thesis, to attempt to rectify these, but rather to explore current means and enable further detail in their application.

3.3.1 A Note on Objective Measurements in TTS

TTS is a rather special case when it comes to evaluation. In most computer science subjects utilising machine learning or modelling, a single definite metric exists with which to evaluate and compare different methods. In ASR, for example, word error rate (WER) provides a single objective measure to evaluate task performance. The lower the WER, the better the system. For part-of-speech (POS) tagging it is the tag accuracy, for spam detection it is the detection accuracy and so on.[2] However, for TTS no such metric exists. Some measurements are used from time to time to demonstrate an improvement, model log likelihood or root mean square error (RMSE) of the pitch, but this is often in the absence of a perceptual result. In fact, these can sometimes be misleading as (Henter et al., 2014) has shown that even with near-perfect models we

[2] Note that even these measurements can be brought into question. For example, it should be clear that some recognition mistakes in ASR are worse than others – if asked "Are you happy?" and one recogniser outputs "Is you happy?" and another "Are you unhappy?" it should be clear that the first is a less harmful mistake than the second, however WER will penalise both words the same.

suffer a perceptual degradation. This is particularly true when we consider the fact that the same sentence can be uttered in several different manners, all being equally valid, e.g. with emphasis placed on different words, and so e.g. RMSE of the pitch could say a produced pitch contour is invalid when it is in fact simply a different valid realisation than whatever realisation tested against. In the face of such problems, even the work on creating automatic instrumental measures of TTS quality is all trying to predict perceptual scores (e.g., Hinterleitner2010, Norrenbrock2012a, Norrenbrock2012, Norrenbrock2015, Soni2016), and although progress has been made, the developed measures are by no means a silver bullet and essentially a tool for development to quickly asses whether a method has merit before the application of a perceptual test to verify the improvement.

As a consequence of this, in this thesis, we will be primarily concerned with subjective evaluation of the produced speech. This is standard (and the standard methods were discussed in the previous section), and, in fact, highly appropriate. The ultimate goal of TTS is to produce artificial speech completely like human speech. That is, our goal is subjective as the only judges as to what sounds human, is human. Thus, with carefully prepared subjective evaluations, such subjective opinion is our main measurement. Objective measurements will be used as a development tool as constant human evaluation is prohibitively expensive, but a perceptual test is always used to validate the objective tool (and sometimes to show the inadequacy of said tool as in Chapter 5). This is not to say that current naturalness tests are perfect, and in fact this is part of the motivation for the work presented in Chapter 4 – to provide subjective tests more robust to opinion by relying on human task performance measure instead of human perceptual preference. Current naturalness tests are, however, the best tools currently available (and the move from mean opinion score test to MUSHRA style tests in later chapters represent a shift toward an improved methodology).

3.3.2 Comparing Read and Spontaneous Speech

A simple way of testing if there is a preference for conversational over read utterances is to adopt the standard naturalness test setup. In such a procedure the common instruction is for the participant to listen to one sentence at a time, rating how natural they find the sentence. That is people are only told to rate what sounds “natural” with no further qualification. If naturalness is an underspecified concept it should be possible to influences people’s ratings by slightly changing the given instructions, and as

Read
Challenge and errors both go well.
Author of the Danger Trail Philip Steel etc.
How funny is your funniest joke?
Officials have no evidence yet that the plane could have been sabotaged.
Spontaneous
It's kinda ridiculous, but it was funny at the time.
When I was younger I... loved uhm Ang Lee.
Absolutely, I'm sure there are evil kings with rotten voices.
And at the point where it goes into the park, the tunnel goes underneath at that point.

Table 3.2: Example sentences of spontaneous and read speech given to participants.

this thesis is concerned with the difference between conversational and read speech it is attempted to influence people's perceptions in either of these directions. Instead of closely matching the content of these sentences by rating the same sentences either spoken in a conversation or read aloud (see Section 3.3.3 for a matched setting), it was decided to initially use sentences representative of the respective styles to see if a difference was to be found in a fairly unconstrained setting.

3.3.2.1 Data

Studio recordings of conversational and read-aloud data from two speakers, one male and one female, were used as the stimuli. These did not come from the above recorded corpus, but rather from proprietary data recorded by Cereproc for their purposes, and kindly lent out for the below investigation. This was necessitated by the fact that this work was performed in conjunction with the recording of the reported corpus. For each speaker, 30 conversational and 30 read sentences were selected. For the read sentences, the female data included mainly read news text and the male data was the first 30 sentences of the Arctic prompts (Kominek and Black, 2003). The conversational utterances were chosen from recordings of the speakers having an unscripted conversation with an experimenter, these recordings were recorded in a similar manner to those of the corpus described above. As mentioned above, all stimuli were recorded by Cereproc for their proprietary use, and they kindly allowed the use of these sentences for this experiment. The sentences were chosen so as to be complete sentences with no initial or final disfluency, although disfluencies were allowed in the sentences. Where the read prompts had a distinct third-person perspective most conversational sentences in the database were first person. To reduce this mismatch, conversational sentences were chosen to generally be about something rather than the speaker him/herself. Sentences in both conditions were also matched for length with the shortest being about 2s long and the longest about 6s. Table 3.2 provides a few example utterances and audio samples are available at the thesis repository besides experimental materials (Dall, 2017).

3.3.2.2 Method

32 paid native speakers of English were recruited, mainly students at the University of Edinburgh. 11 participants rated general naturalness (GenNat), 10 conversational naturalness (ConvNat) and 11 participants reading naturalness (ReadNat). Participants were instructed to rate the sentences in the standard TTS paradigm and they were instructed to “Listen to each sentence and rate it according to how natural you find the sentence from a scale of 1 - Very Unnatural to 5 - Very Natural” in the GenNat case, in the ConvNat the sentence “if you were having a conversation” was added between “sentence” and “from”; in the ReadNat case “if somebody was reading aloud” was added in the same place. This difference in instruction was the only difference between conditions. Each participant rated all 120 sentences once; the order of presentation was randomised for each participant. An additional 5 sentences were run as a trial to allow participants to get accustomed to the methodology. After the trial run, participants were encouraged to ask clarifying questions before proceeding to the main part of the test. Two participants in the general naturalness group enquired what was meant by “naturalness” and the experimenter deliberately gave a vague answer, saying that it was “whatever you find natural”. The test was performed in a soundproof room with the participants wearing good quality headphones. The test took about 15 minutes to complete. There were three groups of participants (GenNat, ConvNat and ReadNat) and two types of audio (conversational or read). In total, the experimental design had six conditions in a three by two design.

3.3.2.3 Results

The 1-5 scale itself has not been much investigated, however the Blizzard Challenge 2008 (Karaiskos et al., 2008) evaluation gave support to the scale being treated, by

	GenNat		**ConvNat**		**ReadNat**	
	Read	**Spontaneous**	**Read**	**Spontaneous**	**Read**	**Spontaneous**
N	660	660	600	600	660	660
Mean	2.98	4.23	2.62	4.04	3.67	3.74
SD	1.192	1.131	1.291	1.189	1.182	1.466
p	p<0.0001		p <0.0001		p=0.352	

Table 3.3: Condition descriptives when comparing read and spontaneous natural speech. N = Number of MOS ratings for condition. Mean = Mean MOS score. SD = Standard deviation of MOS. ConvNat = Conversational Instruction, GenNat = General Instruction, ReadNat = Reading Instruction. The shown significances are between spontaneous and read sentences for each condition.

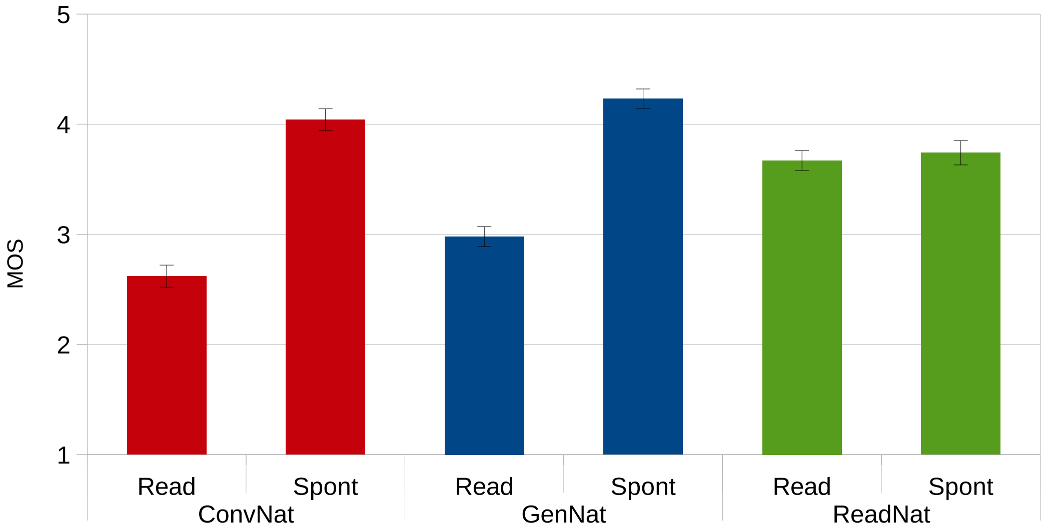
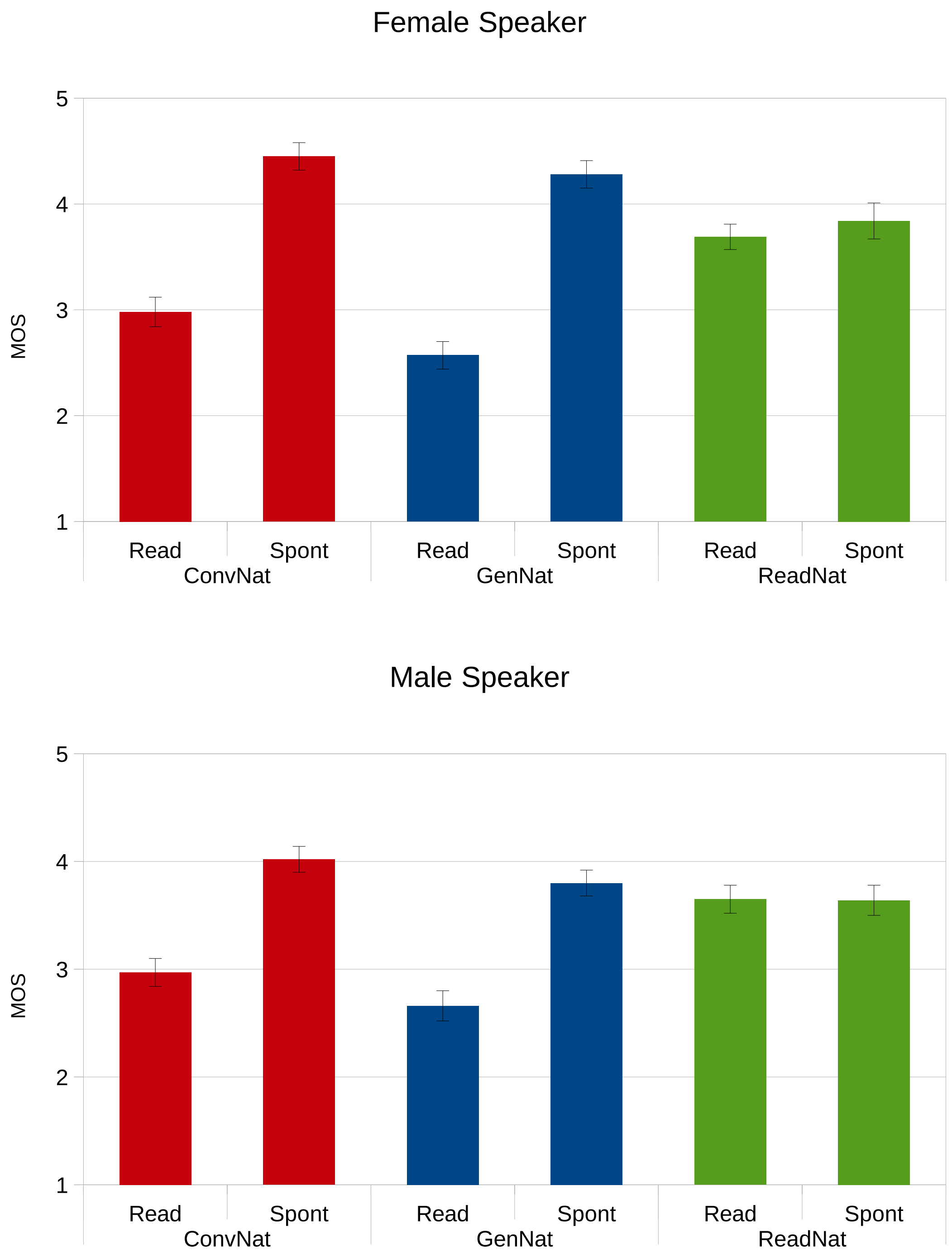


Figure 3.1: Overall MOS naturalness ratings by category. ConvNat = Conversational Instruction, GenNat = General Instruction, ReadNat = Reading Instruction. Spont = Spontaneous.

listeners, as an interval rather than ordinal scale by comparing it to scores obtained using an unnumbered slider. This allows us to analyse MOS-tests by means of standard parametric statistics such as the students t-test, not normally usable on ordinal data. Therefore we can meaningfully compare the means instead of the medians of the ratings (Marcus-Roberts and Roberts, 1987). No null responses were recorded and all ratings were used in the analysis. A significant difference was found between the read (M=2.98, SD=1.192) and conversational (M=4.23, SD=1.131) sentences in the GenNat group (t(1318)=19.644, p <0.001), this was also the case for ConvNat (Read: M=2.62, SD=1.291; Conv: M=4.04, SD=1.189; t(1198)=19.848, p <0.001) but not the ReadNat condition (Read: M=3.67, SD=1.182; Conv: M=3.74, SD=1.466; t(1318)=0.93, p=0.352). In other words, when asked to rate what they found natural with no further instruction, or instructions toward conversation, participants preferred the spontaneous utterances, however there was no such preference when rating naturalness for reading aloud. See Table 3.3. Across the conditions varying instructions, one-way ANOVAs were run for each speech type. An effect for both read (F(2,1917)=122.285, p <0.001) and spontaneous utterances (F(2, 1917)=25.509, p <0.001) was found. After applying Bonferroni correction, all differences were found to be significant at the p <0.001 level for the read speech and for the spontaneous speech all differences were significant at the p <0.001 level except GenNat and ConvNat which was significant at p <0.05. This means that across conditions each type of speech was rated differently according to which instructions were given, suggesting naturalness to be an ill-defined metric which is easy to influence. Informally it can be noted that the two speakers differed greatly in how different they sounded when reading and conversing. The male speaker is an amateur actor and his normal speaking style sounds almost continually acted as he clearly enunciates everything. In contrast, the female speaker differs a great deal in style between reading and conversing - with some participants expressing surprise that it was even the same speaker. It is therefore possible that the findings are speaker specific or gender specific. Looking at the results for each speaker seperately, the effects are slightly smaller for the male speaker and larger for the female, however both speakers exhibit the same tendencies (see Figure 3.2) with the same significant differences suggesting that, at least in this small sample, neither speaker nor gender had an effect.



Female Speaker

Male Speaker

Figure 3.2: MOS naturalness ratings per speaker. Female speaker top and male speaker bottom. Same labelling as Figure 3.1.

3.3.2.4 Discussion

The perception of naturalness changes in the context in which it is rated, by simply adding “if you were having a conversation” or “if somebody was reading aloud” the ratings change. When no instructions were given as to what kind of naturalness to rate, participants find spontaneously produced utterances to be more natural - confirming the assumptions of earlier research that spontaneous speech is perceptually more natural. This is perhaps not surprising, speaking is a more natural way of, well, speaking than reading, but the size of the effect is surprising. It was expected that, when people were instructed to rate according to conversational naturalness, they would have a preference for spontaneous speech, however it is surprising that this effect is similar to that found when asking to simply rate according to naturalness - the total difference between the two types of speech is only 0.17 MOS points larger in the ConvNat condition as opposed to that in the GenNat case. Interestingly in the GenNat case the ratings for both read and conversational speech were significantly *higher* than in the ConvNat case despite the difference between the speech types internally in the conditions being similar. It is not obvious why this might be the case, but one possibility is that in the general case participants are less certain about what they are asked to rate, they have no exact metric to apply and thus end up applying a more lenient metric than in the conversational case where participants are likely to have a clearer idea of what they are rating and this may make them less lenient.

Another surprising result is that in the ReadNat group there was no preference for either mode of speech, that is, even when explicitly asked to rate according to naturalness when reading aloud, participants found spontaneously produced utterances *equally* natural. Thus spontaneously produced utterances are *always* at least as “natural”, if not more, than read aloud speech - strongly suggesting conversational speech to be the, generally speaking, most natural of the two modes of speech.

3.3.3 Separating Acoustics and Text

While we see a difference between the read and conversational speech, the content of the read and conversational sentences was quite different despite ensuring that each spontaneous utterance was “complete”. It is therefore possible that the preferences found are not due to differences in articulation or speech mode - but rather due to differences in content. The opposite, however, is also possible, that is, the content has nothing to say and only the acoustic differences matter. In order to tease this apart further we need to isolate the two options. This can be done in the following way; firstly in order to test whether it is purely the content of the utterance which affect people's perception, we can elicit ratings from people based on text only. That is by comparing normal written text - e.g. from newspapers or novels - with transcriptions of conversational speech we can avoid the acoustic component entirely and focus purely on the content. Secondly we can isolate the acoustic component by recording a speaker in a conversational setting and then, at a later time, ask the same speaker to re-read transcriptions of their own earlier utterances. The content of the utterances will be the same, however the mode of speech will differ. In this way, we can tease apart the effects of content and mode.

3.3.3.1 Data

One acoustic and one textual dataset was obtained. The acoustic data consisted of 50 sentences initially produced in a conversational setting and then re-read by the same voice talent amongst other prompts. This was done in conjunction with the recording of the corpora described in Section 3.2 and it is thus from the same voice talent as those corpora. The 50 sentences were found by transcribing some of the first recordings of spontaneous conversation and identifying 50 sentences which were "complete" utterances as defined above. None of these utterances contained any filled pauses as it was thought that having the voice talent act these would produce decidedly unnatural prompts – unfairly favouring the spontaneously produced prompts. These 50 transcribed utterances were then mixed in with standard prompts for the voice talent to record, without the voice talent being aware that these sentences had earlier been uttered in a conversational setting.

The textual data consisted of 120 sentences. Half were taken from transcriptions of spontaneous data and the other from written sources. The transcribed data was obtained 50/50 from two generally available corpora of spontaneous data – AMI (Carletta, 2007) and Switchboard (Goodfrey et al., 1992). The written data contained 30 sentences from the Arctic (Kominek and Black, 2003) scripts and the last 30 sentences were from news data taken from prompts used in the Edinburgh Voicebank Project (Yamagishi et al., 2012), the same prompt types as used for the read corpus in the previous section. For both types – novels and news, names and quotes were avoided as none were included in the spontaneous data and their length matched the spontaneous in terms of numbers of words. The choice of using various sources for both written and spontaneous data, and the inclusion of disfluencies, was to enable analysis of the

	Examples Sentences
AMI	Yeah, but you can appreciate the way they look.
AMI D	I can address some of that issue, I think, with uh my presentation.
Switchboard	I do try and regulate how much exercise I get a week.
Switchboard D	Yeah that that was a real good one.
Arctic	Unconsciously, our yells and exclamations yielded to this rhythm.
News	The current deployment is designed as a deterrent.

Table 3.4: Example sentences used for comparing different types of text. D = disfluent.

possibility of internal variation depending on the style of the textual data. An example sentence of each type can be found in Table 3.4. Experimental materials and scripts are available at the thesis repository (Dall, 2017).

3.3.3.2 Method

30 paid native speakers of English, mainly students at the University of Edinburgh, were recruited to take part. The general method was similar to the experiment in Section 3.3.2 except as noted below. As before, each participant was assigned one of three groups - general naturalness (GenNat), conversational naturalness (ConvNat) or reading naturalness (ReadNat). The test had two parts. Part 1 consisted of the 50 parallel audio samples and 4 test samples, two spontaneous and two read, drawn from the general read and spontaneous corpus. Part 2 contained the 120 textual samples and additionally 6 test samples, one from each text type. Except fornthe test samples all presentation of stimuli was randomised for each participant. In part 1, participants were asked to rate naturalness according to their group as in Section 3.3.2. In part 2, participants were asked to imagine that the sentence was either “spoken aloud” (GenNat), “said in a conversation” (ConvNat) or “read aloud” (ReadNat), and then judge how natural the sentence would be. In total, the test took about 30 minutes to complete.

3.3.3.3 Audio Results

Of 1500 responses, 15 (1%) were null responses and were excluded. For the GenNat (t(492)=14.864, p <0.0001) and ConvNat (t(496)=10.837, p <0.0001) groups we see a repetition of the previous results with spontaneous speech being significantly preferred over read prompts (Table 3.5). Contrary to earlier we now have a significant difference

	GenNat		**ConvNat**		**ReadNat**	
	Read	**Spontaneous**	**Read**	**Spontaneous**	**Read**	**Spontaneous**
N	248	246	249	249	247	246
Mean	2.79	4.29	2.99	4.09	3.36	4.07
SD	1.292	0.915	1.292	0.938	1.114	1.145
p	*p*<0.0001		*p* <0.0001		*p* <0.0001	

Table 3.5: Condition descriptives when comparing read and spontaneous natural speech when only audio differs. N = Number of MOS ratings for condition. Mean = Mean MOS score. SD = Standard deviation of MOS. ConvNat = Conversational Instruction, GenNat = General Instruction, ReadNat = Reading Instruction. The shown significances are between spontaneous and read sentences for each condition.

for the ReadNat group (t(491)=6.888, p <0.0001) - that is *spontaneous* speech is significantly preferred over read speech (see Figure 3.3). Again one-way ANOVA's were run for each speech type across groups. Here we find that no difference exists for read speech (F(2, 746)=2.693, p=0.068) - ratings of reading naturalness did not change with instructions. However for the spontaneous speech a significant difference was found (F(2, 746)=12.197, p <0.0001) and, after Bonferroni correction, showed the read group to be significantly (at p <0.01) different to the general and conversational group, no difference existed between those (p=0.154). In other words, instructions toward rating for reading naturalness changed peoples perception toward a higher preference for read speech.

3.3.3.4 Text Results

11 responses (0.5%) were null responses and were excluded. In both the GenNat (t(797)=3.877, p <0.001) and ConvNat (t(796)=12.207, p <0.0001) groups the transcribed text was significantly preferred. However, the ReadNat group significantly preferred the *written* text (t(790)=7.694, p <0.0001) (see Table 3.6). When imagining text spoken aloud or said in a conversation, participants found transcriptions of spontaneous speech more natural - but when imagining it read aloud, they found written text more natural. One-way ANOVAs support the hypothesis that instructions affect peoples' perceptions. For the transcriptions (F(2, 1196)=41.058, p <0.0001), the GenNat and ConvNat groups, after Bonferroni correction, differ significantly from the ReadNat group (both at p <0.0001) however not between themselves (p=1). That is, only

	GenNat		**ConvNat**		**ReadNat**	
	Read	**Spontaneous**	**Read**	**Spontaneous**	**Read**	**Spontaneous**
N	399	399	399	400	395	397
Mean	3.36	3.72	2.73	3.81	3.76	3.07
SD	1.385	1.286	1.280	1.213	1.139	1.404
p	p<0.001		p <0.0001		p <0.0001	

Table 3.6: Condition descriptives when comparing read and spontaneous text. N = Number of MOS ratings for condition. Mean = Mean MOS score. SD = Standard deviation of MOS. ConvNat = Conversational Instruction, GenNat = General Instruction, ReadNat = Reading Instruction. The shown significances are between spontaneous and read sentences for each condition.

when rating reading naturalness peoples' ratings are affected by instructions for transcribed text, and then towards being less natural (see Figure 3.4). In the written case there was also a significant effect (F(2, 1196)=58.978, p <0.0001) and with Bonferroni correction all differences were significant (p <0.001). So, when rating written text the instructions consistently affected people's perceptions, people found written text the least natural when rating for ConvNat, more for GenNat and most natural for ReadNat (see Figure 3.4).

As textual domains vary widely, the effect of the text was also investigated. Figure 3.5 shows the mean ratings for each type of text - meetings (AMI), telephone conversations (Switchboard - SB), news (News) and novels (Arctic) - and also the disfluent versions of the transcriptions (AMI-D and SB-D). The overall tendency for both transcribed and written text is clear: transcriptions are preferred when rating conversational naturalness and the opposite for reading naturalness. However, a few points are worth mentioning. Even though, in the ConvNat case, the news text is rated as significantly more natural than the novel text (t(397)=6.029, p <0.0001), all transcribed text is rated significantly higher than the news text (at p <0.005). That is, when rating conversational naturalness people consistently found transcriptions of spontaneous speech more natural than written sources. For the general case, news text was on par with transcribed data, and the novel text was generally least natural, the AMI transcriptions were even considered as natural in the ReadNat case as the novel text (t(296)=0.6635, p=0.508).

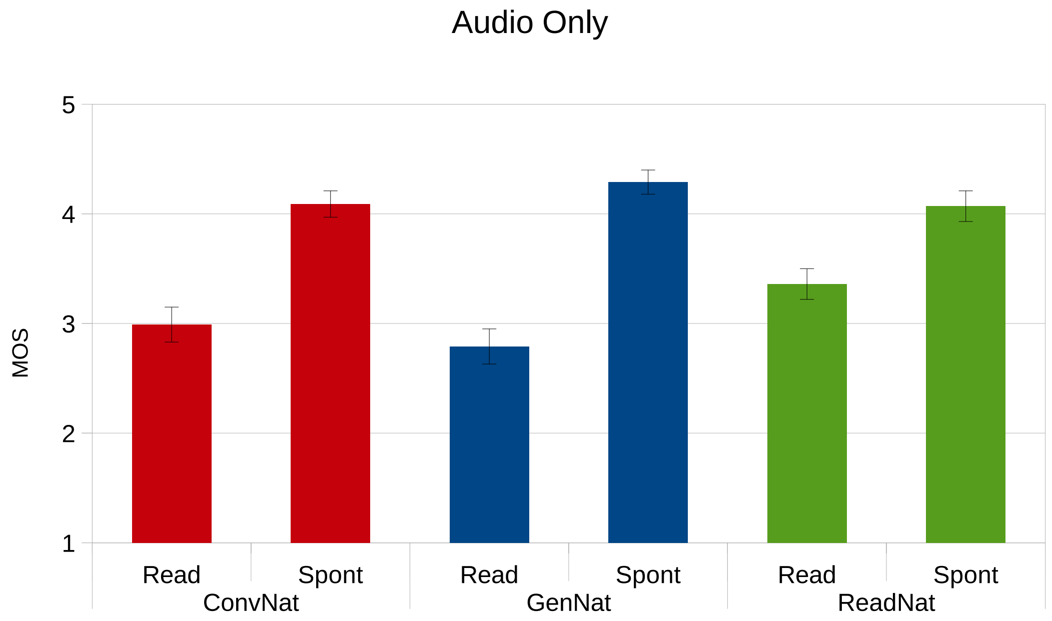


Figure 3.3: MOS naturalness ratings for the content-identical read and spontaneous utterances.

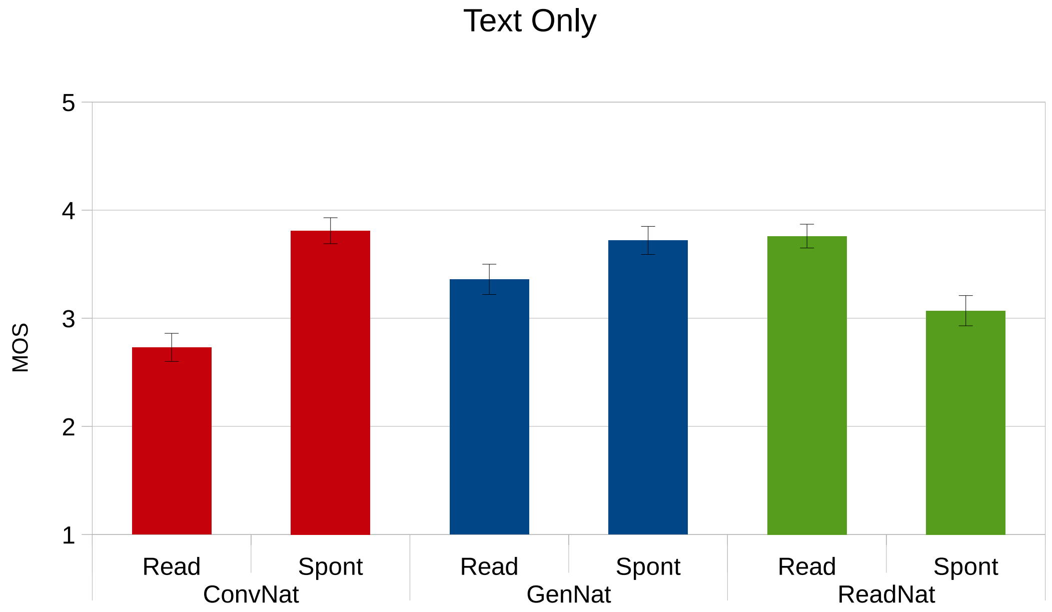


Figure 3.4: MOS naturalness ratings for the comparison between text from written sources or transcribed speech.

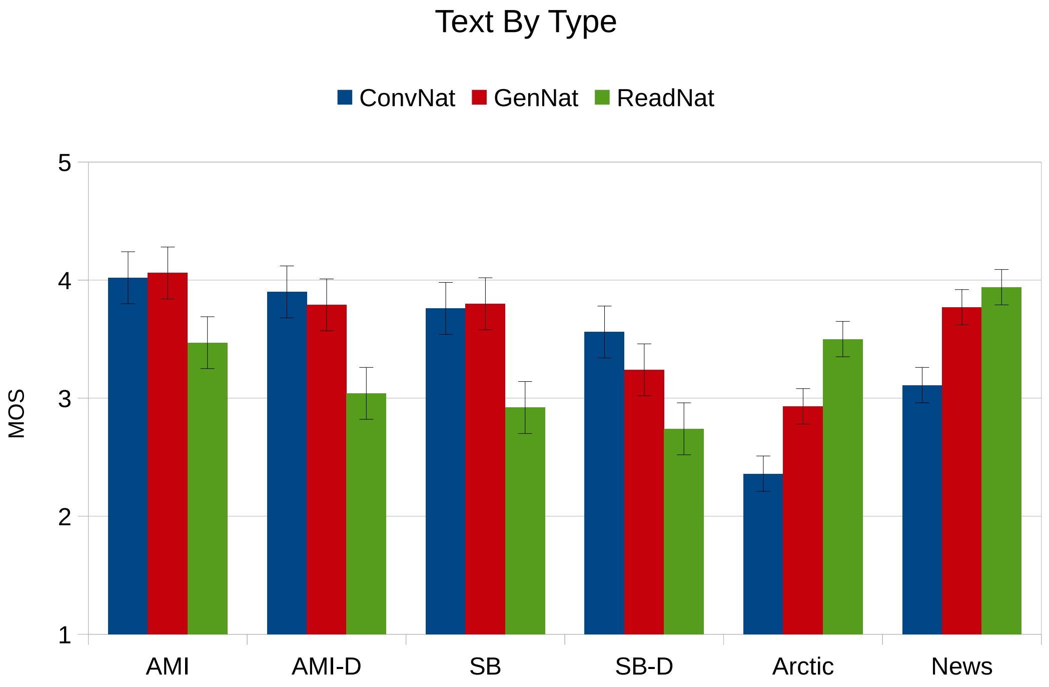


Figure 3.5: MOS naturalness ratings of the text split by type. AMI = Sentences from the AMI corpus without disfluencies. AMI-D = Sentences from the AMI corpus with disfluencies. SB = Sentences from the Switchboard corpus without disfluencies. SB-D = Sentences from the Switchboard corpus with disfluencies. Arctic = Sentences from the Arctic corpus. News = Sentences from news sources of the Edinburgh Voice Bank Project. AMI, AMI-D, SB and SB-D are transcriptions of spontaneous speech. Arctic and News are from written sources, novels and news respectively.

3.3.4 Discussion

The second listening experiment lends further support to the hypothesis that spontaneous conversational speech is considered to be more natural than reading aloud prompts. When considering the audio only, the conversational speech is consistently preferred over the read prompts. This supports the assumptions of earlier studies (Adell et al., 2010a; Andersson et al., 2012; Campbell, 2007) and makes intuitive sense: the spoken word did after all originate in conversation and not in text. The audio results do not line-up exactly with the first experiment in which there was no preference in the ReadNat case. The explanation for this can be found by looking at the overall picture of the textual results, in which we can see a preference for the written sources over the transcribed conversations only for the ReadNat group. Combining the audio and textual results gives us the same picture as was formed in the first experiment (see Figure 3.5 compared to Figure 3.3 and 3.4). That is, we have successfully managed to tease apart the difference between the acoustics and the content of the sentence by removing the variables in their respective tests. It is important to note that, for the textual case, the focus has been on the spoken word, not the written, by instructing participants to rate it according to how natural it would be in various spoken scenarios, and not how natural it would be focusing on it as text. In light of the clear effect of instructions on people's ratings (more below) it would be expected that instructions geared toward *written* naturalness would yield a different result.

Both the first and the second tests support the hypothesis that naturalness as a metric can be easily influenced by experimental instructions, and that the influence is dependent on the type of data under consideration. This is likely due to naturalness being under-specified, and so by conditioning the metric in the experimental setting we can influence our participants towards various interpretations. Knowing this encourages caution and enables more detail when evaluating synthetic speech: caution, because we must be diligent with the instructions we give participants so as not to bias them in an unwanted direction; more detail, as we can condition the metric toward specific aspects of naturalness. For example recent Blizzard challenges have focused on audiobooks (King and Karaiskos, 2012), in such an evaluation we may not wish to evaluate general naturalness but rather what people would expect when listening to an audiobook, as these may be different. As a consequence of this, throughout this thesis, care is taken when giving instructions to participants. While it is tempting to generally steer participants toward a conversational setting, after all that would seem to be the goal of this thesis – more conversational TTS – this was not done. The central hypothesis in this thesis – and the goal – is *not* more conversational TTS at all costs. Rather it is, that by producing more conversational speech, naturalness *in general* will improve. Thus, instructions in this thesis will generally fall into the general naturalness category, and this is justified by the fact that conversational speech is rated as more natural even in the general case. I.e., by producing more conversational TTS we should see an improved general naturalness.

While the audio results are clearly in favour of spontaneous speech in all cases, the picture gets a bit more muddled when looking at the different text styles. There are three things to take from these results. Firstly, while it is possible that the writing style of novels simply is not fit for reading aloud, I believe that it is more likely due to the archaic nature of the out-of-copyright books on which the Arctic script is based. Sentences like "I tell you I am disgusted with this adventure tomfoolery and rot" is far away from everyday language, and here the writing in newspapers is easier to digest. This is of course often an artistic choice by the author of a novel, but it does suggest that we shouldn't rely on such prompts for natural sounding speech. Despite this, this thesis will still use the Arctic set for the read corpus. This is acceptable as although the textual content is odd, we are not synthesising these sentences, which would affect naturalness, but rather we are using them as training data, and the acoustic style of reading aloud is the same regardless of the text content[3]. Secondly, while the overall picture tells us that the transcriptions are preferred in the GenNat case, the news text is actually on par with it, suggesting that at least the textual content can be derived from such prompts. Thirdly, if our focus is on conversational speech synthesis then we are best off focusing on the use of actual spontaneously produced text, which may include disfluencies without important loss in naturalness. Such prompts could be obtained by crawling corpora of transcribed conversations, finding sentences ensuring the usual phonetic balance.

However, as the acoustic results show, we should be better off using actual recorded conversations instead of prompts. These could come from already existing transcribed corpora, but here we encounter the problem of recording quality: very few are recorded under the necessary high quality conditions speech synthesis requires (see Section 3.1). Unfortunately, it is also costly to record and transcribe a corpus of the necessary size,

[3] One can of course imagine scenarios in which somebody reading aloud will change their style drastically, e.g., to put on a voice for a character in a novel. However, for TTS purposes voice talents are instructed to use a neutral reading style and as such this should not change in an important manner regardless of content.

and as Andersson et al. (2012) notes the phone coverage is lower in unscripted conversation and of the recorded data less is directly usable compared to prompts – in the corpus presented in Section 3.2 about 40% of spontaneous utterances had to be discarded compared to only 9 utterances of read.

While all speakers were judged to be more natural when spontaneous in the studies here, it is possible that a sufficiently skilled voice talent can reproduce conversational speech in a prompted situation[4]. Furthermore the voice talents were not instructed to attempt to replicate spontaneous speech but rather their neutral reading voice, which again could result in further divergence. Importantly, however, all read speech samples were representative of current TTS corpora, and as such, the results suggest that current corpora are less natural at the outset compared to a corpus based on conversational speech - supporting the main hypothesis of this thesis.

If this is true it could have profound consequences for how we should be doing speech synthesis. Assuming improved naturalness is the main current challenge in speech synthesis (in particular SPSS) then the results presented here would strongly suggest that we should be utilising the preference for conversational speech by basing our models on such speech. In fact, if we wish to make the most broadly applicable speech synthesis system, we should not assume that read speech is a neutral middle ground, rather that middle ground is more likely conversational speech as shown through the preference for such speech in the general naturalness case.

3.4 Comparing Read and Spontaneous Speech-based Voices

From the previous section we can see that spontaneous speech-based TTS voices have the potential to sound more natural than a voice based on read speech. In contrast, in the previous literature, reviewed in Chapter 2, no improvement was found for spontaneous speech-based voices compared to read speech-based voices. However, Andersson et al. (2012) and Adell et al. (2012) found that by including filled pauses in the speech they achieved a perceived naturalness on par with, and in some cases better than, voices trained on read speech. Therefore, an experiment is here carried out to confirm whether this is also the case for the corpora presented in this chapter. Furthermore, discourse markers were also included. This inclusion was done because of the

[4] Something which will not be investigated in this thesis.

previous experiment showing that word choice also matters, it was hypothesised that a simple way to introduce a more conversational character to the speech through text was to include discourse markers.

3.4.1 Methodology

The read and spontaneous speech corpora used in these experiments were the read arctic-based and spontaneous speech-based corpora described in Section 3.2.

Synthetic voices for both speech types were trained using HTS-2.3beta (Zen et al., 2007a). 60 sentences containing either the FPs ‘UH’ or ‘’UM’ or the discourse markers (DMs) ‘like’, ‘you know’ and ‘I mean’ were extracted from a corpus of spontaneous text consisting of the AMI (Carletta, 2007), Fisher (Cieri et al., 2004) and Switchboard (Goodfrey et al., 1992) corpora and also an unreleased corpus of British conversational telephone speech (see Chapter 9 for details of the corpus). Each sentence was chosen to be a “complete” sentence without the FP or DM and a version of each sentence containing either FPs, DMs or both was synthesised.[5] The FPs were treated as a word in the text stream as in Andersson (2013). This choice is also supported by the findings of Trouvain and Truong (2012).

A 5-point Mean Opinion Score (MOS) test and a preference test to rate naturalness were conducted. For the preference test, the 60 sentences were split into two groups for each of the three conditions (FP, DM or Both) with 30 pairs in each group (for a total of six groups of 30 sentences). For the MOS test the 60 sentences were randomly divided into six groups, each containing 30 read voice and 30 spontaneous voice sentences – ten from each condition (FP, DM or Both). 30 native English speakers were recruited and they performed the test in a sound-proof booth in front of a screen wearing high-quality headphones. Each listener rated one of the two groups of 30 sentences from each of the conditions, presented in a random order, and one of the six MOS groups for a total of 900 preference comparisons and 300 MOS ratings of each condition. The instructions for the preference test asked participants to pick “the sample they found the most natural” without further qualification, and for the MOS test participants were asked to rate for general naturalness as in the previous section. These instructions were chosen as spontaneous speech was rated higher in the general case which was deemed

[5] This was done using the insertion system proposed in Chapter 9 and thus chronologically this test was performed after the work in Chapter 4 and much of the work in Chapter 5. It is however presented here and not in Chapter 9 as it fits better into the thesis structure. Note that the used system did not employ any of the methods proposed in this thesis and is as such a standard system.

	Read	Spontaneous	No Preference
Filled Pause	52.5%	31.6%	15.9%
Discourse Markers	54.1%	32.9%	13.0%
Both	52.1%	33.0%	14.9%

Table 3.7: Preference scores between the read and spontaneous speech-based voices synthesising sentences with FPs, DMs or both.

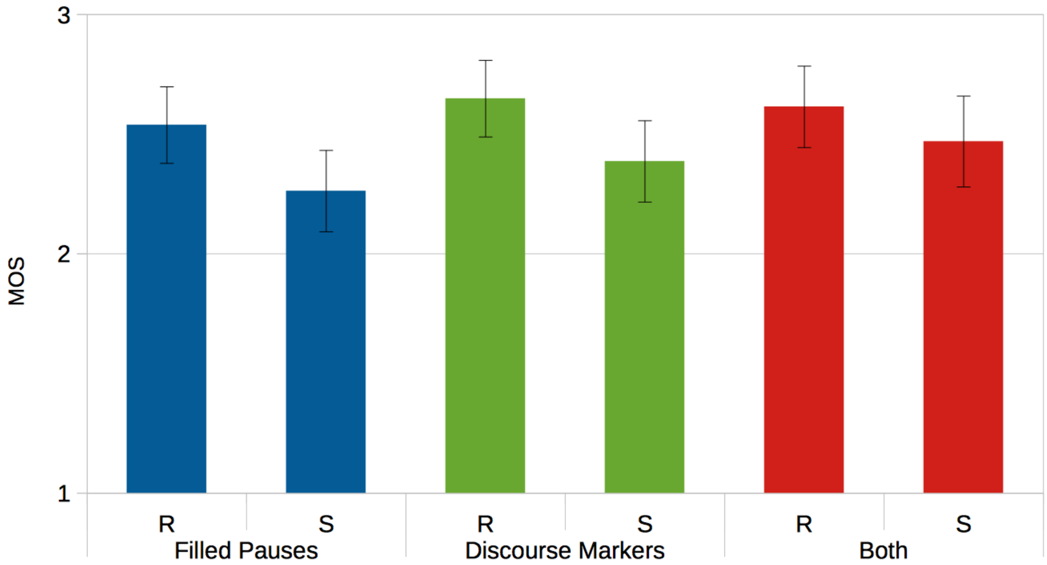


Figure 3.6: MOS ratings between the read (R) and spontaneous (S) speech-based voices synthesising sentences with FPs, DMs or both.Y-axis scaled from 1-5 to 1-3 for clarity.

more generally useful than the instructions steered toward conversational naturalness. Experimental materials available at the thesis repository Dall (2017).

3.4.2 Results and Discussion

Due to one participant misunderstanding the instructions for the experiment, this participant's results were not included in the analysis. All no preference scores were pooled to allow for statistical analysis using the exact binomial test. Table 3.7 shows the preference test results. Participants preferred the read speech-based voice over the spontaneous speech-based voice in all cases and this was significant, after applying Bonferroni correction, for all conditions (at least at $p < 0.001$). For the MOS test, the read speech-based voice was rated significantly higher than the spontaneous in all cases (Figure 3.6). This difference was significant for the FP (t(288)=2.32, p<0.05) and the DM (t(288)=2.21, p<0.05) case but not for the Both case (t(288)=1.11, p=0.27). In general, these results shows that the read voice was considered more natural than the spontaneous. The difference when including both FPs and DMs in the MOS was smaller than for the other conditions, however it was the same in the preference test and is probably a results of a low number of ratings. Suggesting that the straight inclusion of DMs in the speech does not improve naturalness. These overall results are in contrast to the findings of Andersson (2013), Andersson et al. (2012) and Adell et al. (2012). These earlier experiments found that a spontaneous synthetic voice trained on data that contained FPs was rated at least as natural as a voice trained on read speech. There are, however, some differences in the experiments performed. In Andersson's work (Andersson, 2013; Andersson et al., 2012), a data-mixing technique was applied to overcome data sparsity problems, in which one voice was trained using both types of speech, but a linguistic feature was added marking each sentence with the source speech type. At synthesis time, sentences could then be synthesised with either tag, and this technique enabled better FP-containing sentences to be synthesised by means of the spontaneous tag. In Adell et al. (2006), synthesis of the FPs was based on a specific FP model, which was subsequently improved in Adell et al. (2010a). FPs were modelled separately from other speech by applying modified search rules in a unit selection system. However, in both approaches, no results are given of the quality of synthesis based only on the data containing FPs. This suggests that their methods must be responsible for closing the gap in naturalness that was found here between synthetic speech-based on read speech versus that based on spontaneous speech – this will be investigated in Chapter 7.

Looking at some of the produced speech some things are worth noticing. Firstly, listening to the samples it is clear that there are prosodic issues with the produced spontaneous speech, this is, however, clearly also the case for the read speech (e.g. sample 29 of the MOS sentences with both DMs and FPs in the relevant part of the thesis repository - Dall (2017)). What is notable is that the errors made by the spontaneous voice are more serious when they happen. A good example is some of the FPs produced. In Figure 3.7 the spectogram and waveform of the final “but UM” in sample 14 of the MOS sentences with both DMs and FPs is shown. What is notable is, that although the spontaneous voice generally produces faster speech, the UM is indeed longer than that produced by the Read voice as would be expected, however it is very buzzy with a particularly noisy tail. This kind of issue can be likened to joining mistakes in unit selection synthesis, it doesn't matter how good the rest of the sentence

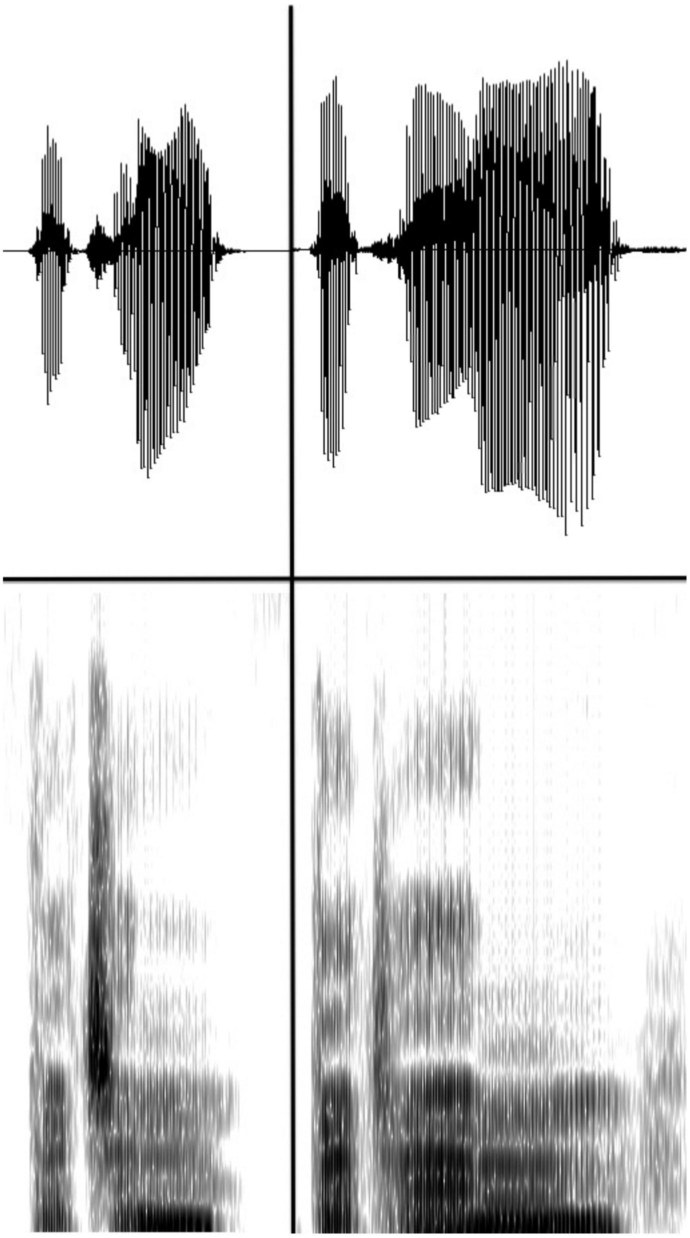


Figure 3.7: Waveform and spectrogram of “but UM” produced by the Read speech based voice (left) and the spontaneous speech based voice (right).

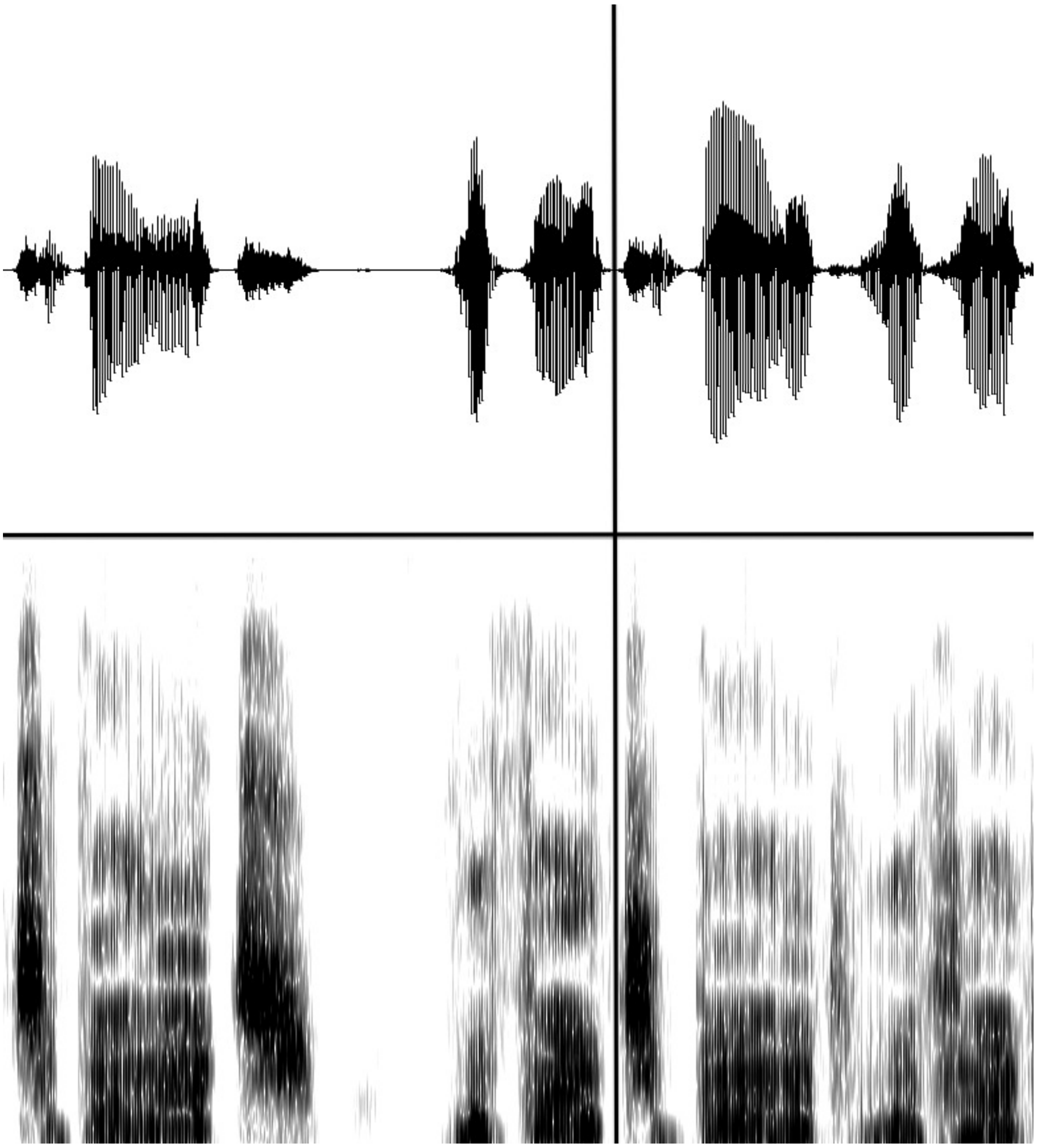


Figure 3.8: Waveform and spectrogram of “to buy it without” produced by the Read speech based voice (left) and the spontaneous speech based voice (right).

might be, people won't like it if such an error exists. In fact, the main perceptual issue with the spontaneous voice is a certain amount of choppiness. Sample 1 of the MOS sentences with both DMs and FPs illustrates this very well, Figure 3.8 shows a section of that sample (the reader is encouraged to listen to the full sample in the thesis repository) illustrating the issue. Two things are notable in the highlighted sample, firstly the choppiness is evident when looking at the spontaneous sample compared to the read in which each word (or compound) is much more uniform in duration and energy, whereas there is more variation evident in the read sample. The second is that the predicted pause in the read speech is non-existent in the spontaneous, despite this being in the labels.

Figure 3.9 shows a natural sample (part of 004_3 from the spontaneous corpus) showing how smooth the natural speech is compared to that produced by the synthetic voices, where there are clear closures for the "without" which are not present in the natural speech (this is in fact the case for all "without"s in the spontaneous corpus). This suggests that there are issues with the general modelling of the spontaneous speech and not just with the prosodic realisation. In fact, bad prosody is commonly referred to as one of the main issues with Read speech based voices, and since these issues seem to be related to general modelling and not just prosody, the focus in Chapters 5 and 6 is on pronunciation variation, particularly reduced pronunciations, as a potential explanation for this bad modelling as the missing pause and choppiness suggests that individual phones are not well modelled.

3.5 Chapter Conclusions

In this chapter, I have presented a methodology for recording a corpus of spontaneous speech suitable for TTS based on the method described by Andersson (2013). I have presented such a corpus, recorded together with a parallel standard read speech corpus from the same voice talent. As the read corpus is larger than the conversational corpus, it has been reduced in size to better approximate the size of the spontaneous one. This was done by taking the part based on the Arctic set of prompts. Furthermore, an investigation into the naturalness of read and spontaneous speech was presented. I have shown that MOS-scale ratings can be employed to distinguish the conversationality of speech, in fact spontaneous conversational speech is found to be more natural by listeners than read prompts. We can also affect people's perception of naturalness by simple conditions in the instructions, enabling greater control over the testing scenario

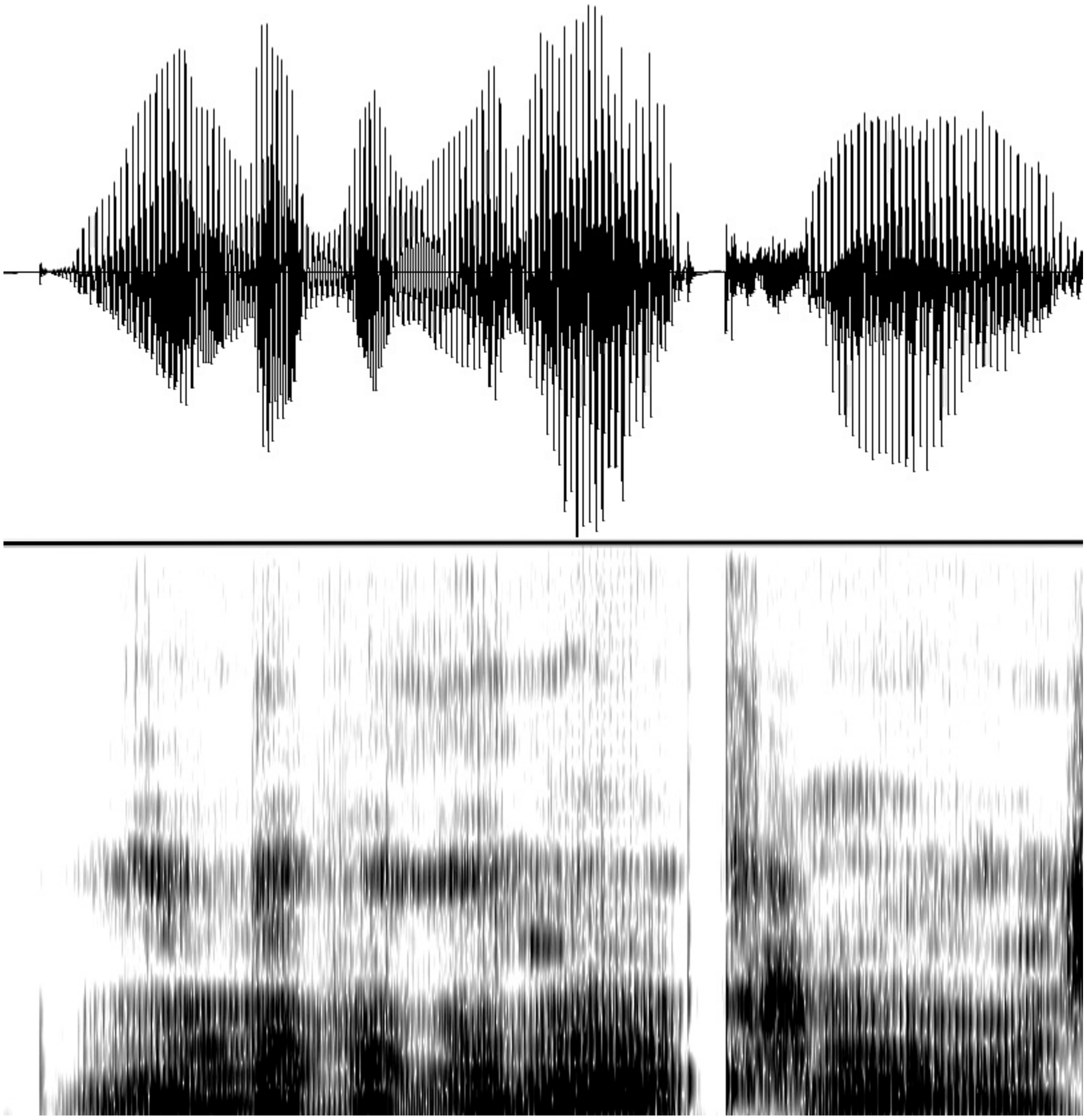


Figure 3.9: Waveform and spectrogram of “well living without a car” from sentence 004_3 of the spontaneous corpus.

while also cautioning its use.

In support of the main hypothesis of this thesis, the results suggests that using read prompts as the basis for TTS may not be producing the neutral general speech previously assumed and that this role is more likely attributable to spontaneous conversational speech. However, the comparison of a voice based on read and a voice based on spontaneous speech presented in the end show that spontaneous speech is not modelled well by current TTS methods.

Chapter 4

The Subconscious Effects of Filled Pauses in Synthetic Speech

A Note on Collaborative Work

The work presented here has been done in close collaboration with Martin Corley (School of Philosophy, Psychology and Language Sciences, University of Edinburgh) and Mirjam Wester (CSTR). For the reaction time-oriented section, both were invaluable partners for general discussions, Martin in particular with regards to statistical analysis and Mirjam with regards to stimulus preparation and manipulation. Catherine Lai was the speech expert who checked the stimuli for noticeable edits and Marcus Tomalin provided the DARPA EARS data analysis. The reaction time work was previously published as Dall et al. (2014b) and is here presented in rewritten form. For the Change Detection experiment, Mirjam provided valuable discussion and insights into the experimental paradigm, Martin provided the scripts and natural sentence data used for the test and two Masters students, Amelie Osterrieth and Anisa Jamal, collected the data for the natural speech, however subsequent analysis and experiments were carried out by myself. The change detection experiment has been published as Dall et al. (2015) and is here presented in rewritten form. Furthermore additional notes are drawn from Wester et al. (2015b) and Wester et al. (2015a) for which Mirjam did the main work and I participated as a discussion partner.

4.1 Introduction

In the previous chapter, an investigation into the naturalness of read and spontaneous speech was performed, in which it was shown that spontaneous speech is perceived to be more natural than read speech. Naturalness, however, is only one goal of this thesis; another is the use of spontaneous conversational phenomena.

One pertinent type of phenomenon is disfluencies. Broadly speaking, disfluencies can be defined as the non-propositional content of a sentence (Shriberg, 2001; Fox Tree, 1995). In this broad definition a "disfluency" includes discourse markers (e.g., 'you know', 'like', 'I mean', etc.), filled pauses (e.g., 'UH', 'UM', etc.), repetitions (e.g., 'it it', 'the man the man', etc.), restarts (e.g., 'the man- manager', 'it was go- it was great', etc.) and affective sounds (e.g., laughter). In this chapter, the focus will be on the category of filled pauses, under which it will be argued that repetitions and some discourse markers belong. Discourse markers are arguably not "disfluent", however they can be treated in a manner similar to that of filled pauses. We will see how in Section 4.6 and repetitions likewise, as will be shown in Section 4.3. Although affective sounds constitute an interesting element of conversational speech, they fall outside the current scope, however work can be found on, e.g. laughter, which the interested reader can find in Campbell (2006a); Trouvain and Schröder (2004); Cakmak (2016). Restarts also fall outside the scope of this thesis, they are a much harder problem to tackle, with less obvious potential benefits to their inclusion. One such benefit could be in incremental systems, however, the system of e.g., Baumann and Schlangen (2012, 2013) and Baumann (2013) does not employ mid-word restarts, but rather a hesitation strategy or simply a switch in content when necessary - a word which is being uttered at the time of content change is finished first.

Filled pauses (FPs) are non-silent pauses in speech and generally this is considered to include 'UH', 'UM', 'UHU', 'UHUM' and 'MHM'[1], but utterances such as 'OH', 'AH' and word repetitions have also been similarly classified (Fox Tree, 1995; Fox Tree and Schrock, 1999; Adell et al., 2006). FPs can occur as part of a sentence or as backchannels. In this chapter, the focus will be primarily on the effects of mid-sentential FPs due to the nature of the experimental paradigms, however, please see Chapter 7 and 9 for the development and use of synthetic FPs in all parts of a sentence. This thesis will not be considering backchannels as this is arguably a separate

[1] Spellings differ and some common alternatives are 'ER', 'EM', 'UHM' and similar. In Chapter 3, in the description of the recording methodology, many other alternatives are described.

phenomenon – the interested reader can have a look at the work by Oertel et al. (2016).

As FPs have, from a traditional linguistic perspective, been seen as disfluencies and thus carrying no content, they have been considered something which could be removed from speech without consequence (Shriberg, 1996). Despite this, they are extremely common in speech, estimates range from 2% to 26% (Fox Tree, 1995; Shriberg, 1999; Bortfeld et al., 2001) - though 6% is often cited (Fox Tree, 1995) and this is probably closer to the truth as studies reporting higher figures often also include other phenomena such as restarts and discourse markers. Corpus studies have described a number of general acoustic properties of FPs and in Section 4.2 some of these are presented with an analysis of the collected corpus of spontaneous speech from the previous chapter. Furthermore, psycholinguistic experiments have shown FPs to carry a number of benefits to the listener. In a series of experiments Fox Tree has shown that reaction time (RT) to a target word following an FP decreases compared to a pause of equal length or complete omission. RT was measured by having participants presented with a word in text, the target word, they would then listen to sentences of speech and were asked to press a button as soon as they heard the target, RT was then measured as the time, in milliseconds, from target word onset to button press. The manipulated variable was then whether the target word was preceded by a normal silent pause (SP), no pause or an FP. It was shown that the presence of an FP made participants react 30-50ms *faster* to a target word than when the target word was preceded by an SP or no pause. This was shown for 'UH's (Fox Tree, 2001), 'OH's (Fox Tree and Schrock, 1999) and repetitions (Fox Tree, 1995) in English, was also found in Dutch (Fox Tree, 2001) and was confirmed by Corley and Hartsuiker (2003). This also constitutes evidence for the similarity between FPs and repetitions. While these reaction time experiments provide evidence that FPs, and also repetitions, affect peoples' online processing, FPs may have other, and longer term, effects. Change Detection Sturt et al. (2004) is a paradigm in which participants are asked to listen to short paragraphs of speech and are subsequently presented with the contents of the speech in writing. It is then the task of the participant to detect if a single change has occurred in the text compared to the speech. This requires participants to not only process the speech as it is heard, but also to memorise it long enough to detect a change at a later point. Thus change detection, as opposed to reaction time, experiments provide a measure of the memorability of the speech in a slightly longer term context. In Chapters 6 & 7 Collard (2009) reports that the presence of an FP prior to the changing word, as compared to fluent speech, increases the change detection rate by 10-15%, and concludes that the acoustic quality of the FP is responsible for the effect (Chapter 7.6, pp. 128). Additionally, other benefits of FPs have been shown, including better recall of target words (Fox Tree and Schrock, 1999; Corley et al., 2007), identifying target objects more accurately (Brennan, 2001) and easier integration of words in their contexts (Corley and Hartsuiker, 2003; Corley et al., 2007). They have also been shown to lead to a better impression of a speaker when not specifically attended to (Christenfeld, 1995), can affect the perception of a speakers personality in natural (Laserna et al., 2014) and synthetic (Wester et al., 2015a) speech and when used as a delay strategy they can improve the perception of both dialogue systems (Baumann and Schlangen, 2012, 2013; Baumann, 2013) and robot interactions (Shiwa et al., 2009, 2008).

The above studies show how FPs play important roles in conversational speech and serves as motivation for the focus on (this chapter) and modelling of (see Chapters 7 and 9) FPs in TTS. If FPs can be integrated into the TTS system in a natural way we should be able to observe the same benefits as those shown in the psycholinguistic studies of reaction time, change detection and the like - generally improving the understanding and intelligibility of TTS in subtle useful ways. Whether these effects are already observable, and thus whether current TTS systems can already convincingly produce FPs, is explored in this chapter. Furthermore, the studies of (Shiwa et al., 2009, 2008), in which robot interactions were rated higher when the robots utilised (pre-recorded natural) FPs, provide evidence that the use of FPs can help humans interact with life-like agents, and due to the general rising use of both robotic and avatar-based life-like agents it is of increasing interest to explore means by which FPs can be used in TTS.

The aim of this chapter is to evaluate FPs produced by current standard synthesis systems, using standard read speech, in the context of such psycholinguistic research, specifically by repeating the reaction time studies of Fox Tree and the change detection studies of Collard in the context of vocoded and synthetic speech. Vocoding was included in the experiments, as, in speech synthesis, it is the step of parametrising the speech in a manner suitable for statistical machine learning. This parametrised version can be re-formed directly by the vocoder, with some loss in quality, and this vocoded speech is thus a form of upper bound on the achievable quality of synthetic speech. Consequently if the previously found effects of FPs do not appear in vocoded speech we can identify the vocoding step as a limiting factor in the realisation of FPs.

In general, FPs are not considered in speech synthesis systems. To date, few attempts have been made at modelling and inserting FPs. Previous studies by Adell and

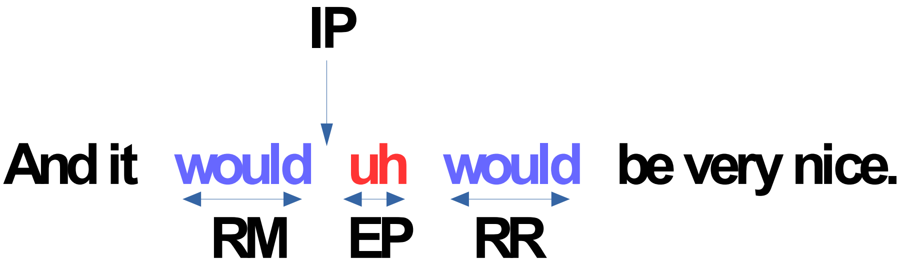


Figure 4.1: The description of a disfluency by Shriberg (1994) and Levelt (1983). RM = reparandum, IP = interruption point, EP = editing phase and RR = repair.

colleagues (Adell et al., 2007a, 2010a, 2012) included FPs in concatenative speech synthesis by constructing a specific 'disfluency' model in unit selection synthesis using what they called the underlying fluent sentence (Adell et al., 2012). This model is based on that of Shriberg (1994) (which is in turn based on Levelt (1983)) which defines a disfluency using the idea of a reparandum, interruption point, editing phase and repair – see Figure 4.1. In this model, disfluencies can be categorised according to their effect, an FP thus occurs in the editing phase (EP) and may or may not have a repair (RR) following it. Based on an analysis of the effects of an FP on its containing units and the surrounding units, Adell and colleagues developed a number of specific target costs for disfluency modelling and applied those.

Another approach (Andersson et al., 2010b, 2012; Andersson, 2013), which uses Hidden Markov Model (HMM) synthesis, treats FPs as normal word tokens in the speech stream when building models based on spontaneous speech. Andersson et al. (2012) presents a method for mixing spontaneous and read speech recordings (see Chapter 7 for details), which, when synthesising with a conversational setting provided improved FP synthesis. Andersson et al. (2012) showed improvements in perceived conversationality and Adell et al. (2012) showed users prefer a system which includes FPs, and both report naturalness matching that of state-of-the-art systems based on read speech. However, the evaluation of the effect of FPs in synthetic speech is, unfortunately, not entirely convincing. For example, the evaluation in Adell et al. (2007a) consisted of comparing pairs of sentences with/without FPs and asking questions specifically regarding FPs (e.g. ”Do you think that filled pauses make the voice (more/equal/less) suitable for a dialogue?”). The perceptual results supported what the authors were hoping to find, i.e., sentences with FPs were judged to be more natural, equally suitable for dialogue, and more humanlike. Considering the investigation in the previous chapter, this questioning would have the effect of priming participants toward favouring the FPs, disregarding other potential issues with the voices such as reduced overall quality (see Chapter 7 for a detailed discussion of this).

In this chapter, the investigations take their starting point from standard read speech-based voices and not the conversational speech recorded in the previous chapter. There is a practical and a theoretical reason for this. While preparing these experiments the database of spontaneous speech was not yet fully transcribed and ready for use in TTS, effectively preventing the use of spontaneous speech for the experiments. However, while spontaneous speech would have allowed the use of e.g., Andersson's approach, this approach is non-standard and cannot readily be applied without special data. It is of interest to see how a standard TTS voice performs in these tasks before attempting to apply more specialised voices, because if the standard voice already works, then there is no point in trying to "fix" it. See Chapter 10 for a revisit of these psycholinguistic studies in the context of the methods for better FP synthesis developed in this thesis.

In Section 4.3, an investigation based on the findings of Fox Tree (1995); Fox Tree and Schrock (1999); Fox Tree (2001) is presented, in which her methodology is applied to FPs from vocoded and synthetic speech in addition to recordings of natural speech. Following that, in Section 4.6, a similar investigation is performed in which the use of vocoded and synthetic speech is used in a change detection experiment.

4.2 Acoustic Effects of Filled Pauses

When considering FPs we can note certain regularities, both in use and in realisation. In terms of their use, corpus studies suggest that they tend to appear around phrase boundaries and before multi-syllabic words (Shriberg, 1994, 1996; Blackmer and Mitton, 1991) and as noted above they are extremely common in speech, representing roughly 6% of speech tokens.

Studies have shown that the fundamental frequency (F0) contours and the duration of FPs are different to other phones in fluent contexts (O'Shaughnessy, 1992; Shriberg, 2001; Adell et al., 2012) – this is primarily found to be a lowering of the F0 and an increased duration. Another characteristic of FPs is the presence of silence before and/or after the filled pause (Clark and Fox, 2002; Adell et al., 2012).

Adell et al. (2012) found, in their corpus, that a preceding silence often occurs (60%) but following silence only occurs in 24% of cases, though they noted that this was possibly an anomalous finding due to other studies reported higher rates. Other papers, which explore these phenomena in different languages and different kinds of corpora, report varying patterns of silence after FPs.

To verify these studies an analysis of the corpora presented in the previous chapter was done. An analysis of the duration and F0 values of ‘UH’ and ‘UM’ was carried out, including their immediately preceding and following syllables, as well as the overall phone statistics of vowels and consonants in the corpus.[2] This was also done for the read corpus, while it does not contain any FPs it is an interesting point of comparison to some of the effects of speech mode noted in Section 2.2.1.

If we look at the details in Table 4.1, we can note that the read speech, in general, has slightly longer in duration, particularly for silence (though the medians tell a slightly different story), but perhaps not as long as one would expect considering the general speaking rate effect found previously.[3] A lower F0 for the spontaneous speech is also notable compared to the read, something this author has not seen in previous research. For both ‘UH’ and ‘UM’ we can note that their F0 is close to that found for spontaneous speech in general. However, looking at the vowel of the immediately preceding or following syllable we can note a general fall in F0 for the FP as is expected from previous research, this is particularly evident when looking at the mean difference between the FP in its left or right context where there is a fall in all cases, also for the ‘UM’ in which the mean does not tell the same story. This difference is explained by the fact that 74.1% of ‘UH’s were preceded by a silence, with 42.6% followed by a silence and 75.5% of ‘UM’s were preceded by silence and 68.9% were followed by silence. In those contexts we cannot calculate the surrounding F0 and thus that makes the overall mean less meaningful. For the silences we can note that the left silences are generally similar to those found in natural speech, whereas the right silences have a much higher median length. For ‘UH’ the realisation is a single vowel which is much longer than the average vowel of the spontaneous speech – for the ‘UM’ it is less easy to compare but the duration is clearly longer than one average vowel and one average voiced consonant (i.e., the phones of an ‘UM’). These findings are all in line with previous research.

So what we can say is that FPs have *longer* durations than normal phones, *lower* F0 than their immediate context and are often accompanied by silent pauses.

[2] No attempt at matching the specific phone type was done as we will see in Chapter 7 a specific separate phone model is usefully employed.

[3] These numbers are derived from the Combilex pronunciation variant forced alignment presented in Chapter 5. Which may help to explain why the duration seems less different than expected as there are in general less phones in the spontaneous.

	Dur mean	**Dur median**	**F0 mean**	**F0 median**
Spontaneous				
Vowels	61.2	46	172.5	169.5
Consonants	80.5	68	171.4	168.0
Silence	165.3	102	-	-
Read				
Vowels	67.6	54	188.2	185.9
Consonants	81.4	78	184.8	181.8
Silence	496.0	80	-	-
UH				
UH	221.1	214	174.5	175.1
Left Syll Vowel	99.4	70	177.8 (9.3)	180.0
Right Syll Vowel	67.6	60	180.0 (12.8)	178.5
Left Silence	205.4	90	-	-
Right Silence	211.1	158	-	-
UM				
UM	373.8	360	170.3	167.5
Left Syll Vowel	-	-	184.8 (23.7)	181.8
Right Syll Vowel	-	-	170.1 (23.2)	173.4
Left Silence	199.8	90	-	-
Right Silence	228.4	190	-	-

Table 4.1: Details of phone durations and F0 for the read and spontaneous corpus and for ‘UH’ and ‘UM’ and their contexts. Dur = duration. Syll = Syllable. Durations are in ms and F0 in hertz. Durations around UM left out due to not being comparable. Silence has no F0. Consonant F0 is for voiced consonants. The number after the slash for left and right contexts is the mean deviation from the FP in context.

4.3 The Effect of Filled Pauses on Reaction Time

For this study, the method of Fox Tree (1995); Fox Tree and Schrock (1999) and Fox Tree (2001) is followed with a few small adjustments - most notably the use of vocoded and synthetic speech in addition to natural speech. The general method involves visually presenting a target word to listeners and asking them to react as quickly as possible when they hear the target word, by pushing a button. In Fox Tree's experiments, each stimulus was presented in three conditions: i) FP included, ii) FP replaced by a silent pause (SP) of equal length or iii) FP spliced out of the sentence[4]. In the current experiments, the first two of Fox Tree's conditions are included: the FP and silent pause of equal length (SP) conditions. Having no pause at all was omitted because it is more interesting whether information may be understood faster with FPs than without, and even if the effect appears when there is no pause, the non-existence of a pause reduces the sentence length more than the magnitude of the previously found FP effect. It has also been argued by Corley and Hartsuiker (2011) that there may be issues with preceding and trailing silences in Fox-Tree's third condition (of having no pause), and finally, leaving out the no pause condition reduces the size of the experiment to something feasible when also considering two additional forms of speech. Furthermore, here all three of 'UH', 'OH' and repetitions found by Fox Tree to have the RT effect will be treated as one and included in the experiment. While repetitions are to some extent not classical FPs, Adell et al. (2006) show that they exhibit similar patterns. As such it is fair to treat them in the same context as other FPs.

Participants hear only one version of each sentence, and these critical stimuli are interspersed with filler sentences. The filler sentences are sentences which do not test any of the conditions in the experiment, but which are included in order to avoid bias effects in which participants may subconsciously pick up on the FP or SP as a cue to an upcoming target word. These filler sentences are not to be confused with sentences containing FPs before a target word which are considered critical stimuli. In the following discussions, filler sentences or stimuli should thus be considered in relation to critical stimuli and not whether or not the sentence contains FPs or SPs, which will be refer to as the FP or SP condition/sentences respectively. The critical measurement of the experiment is the participants' RT to the target word. The target word is defined as the first non-determiner after the FP as in Fox Tree (1995), see Fox

[4] Note that Fox Tree did not always replace the FP with an SP or splice the FP out, she also tested if splicing it in, or replacing an SP with an FP had the same effect. As it did, this is not a potentially confounding factor.

Tree (2001); Fox Tree and Schrock (1999) for variants. This methodology was applied using naturally occurring speech, vocoded and synthetic speech.

4.3.1 Data

116 utterances containing an instance of either 'UH', 'OH' or a repetition, all referred to as FP here, were selected from the AMI Meeting corpus (Carletta, 2007), a corpus of spontaneous speech from meetings. Care was taken that the FP in these utterances was followed by a word that did not appear earlier on in the sentence. There were 79 critical stimuli and 37 filler stimuli from 5 differing speakers (three male and two female). For the critical stimuli, the target word was preceded by an FP and for the filler stimuli the target word was not preceded by an FP. These filler stimuli were included to avoid listeners being primed by the presence of the FP due to the experimental setup rather than due to any effect of the FP itself. The natural stimuli were digitally edited such that the FP was replaced with a silent pause (SP). Note that silent does not mean silence, but rather that a pause (not containing speech) was taken from another point in the same recordings and edited in. This was done, instead of editing in complete silence, to ensure that the SP had similar background noise as the rest of the sentence. A speech expert was asked to check the data for edits but was unable to identify them reliably. Next, the utterances were vocoded using STRAIGHT (Kawahara et al., 1999b), a state-of-the-art vocoding technique. The HMM-synthesis voice was based on HTS 2 (Zen et al., 2007a) in a system that was newer than, but broadly similar to, that in Yamagishi and Watts (2010), which is representative of the state-of-the-art and was based on about 8 hours of training data from a British English female speaker.[5]

During synthesis, the FPs were treated as regular word tokens in the input stream, as argued for in Clark and Fox (2002), which was also the case for silent pauses (SPs) and ensures both are treated the same by the system. When producing the SP versions of the sentences the FPs were not edited out, but rather replaced with a pause in the input specification (the full-context labels), and the same durations were enforced when synthesising. Due to the nature of HMM-synthesis, the exact durations of filled and silent pauses can deviate by a few frames (frame = 5 ms). This deviation occurs due to slightly differing paths through the HMMs in either case as HTS does not unilaterally follow input timings when asked. Although there are slight differences in timings, this was deemed the better solution as it avoids manual editing of the synthesis output and

[5] The read speech corpus presented in the previous chapter was not utilised for this due to the work being carried out in parallel.

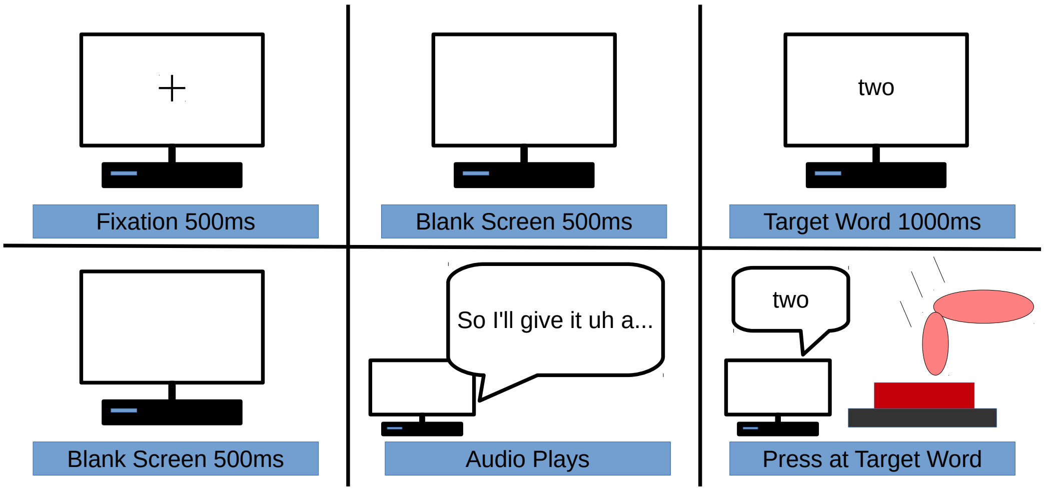


Figure 4.2: An illustration of the RT experiment described in Section 4.3.2.

better modelling of, e.g., co-articulation. All used stimuli and experimental materials are available at the thesis repository (Dall, 2017).

4.3.2 Method

Thirty native English speakers (mainly students at the University of Edinburgh) participated in this experiment. The test took approximately 25 minutes and participants were paid. Each participant was seated in front of a computer screen in a sound-attenuated booth. A fixation point was presented visually on screen for 500 ms, this was followed by a blank screen for 500 ms, and then by the target word for 1000 ms. 500 ms after the target word disappeared, an utterance was played and participants were instructed to press a button as quickly as possible if they heard the target word. Participants were instructed to only press the button if they heard the word. Figure The test was split into four parts. The first part was a trial run; this always consisted of four stimuli, one each of the natural, vocoded and synthetic filler stimuli and one critical stimuli from one of the three speech types. Participants were encouraged to ask clarifying questions after the trial run. The remainder of the stimuli – 34 filler and 78 critical of each speech type – were divided into three roughly equal sized parts from each speech type and randomised. All participants were presented with each utterance only once in any of its forms (with an equal amount of each form). In total, the experiment consisted of six conditions, two versions (SP and FP) of each of the three types of speech (natural, vocoded and synthetic).

4.3.3 Results

Due to experimenter error, three synthetic sentences were wrongly synthesised and excluded from the analysis. Of the remaining 2305 critical responses, 141 were null responses (where participants did not press the button) and excluded. Outliers were determined using the median absolute deviation (MAD) of Leys et al. (2013) instead of standard deviation (SD), because SD is itself subject to outliers, MAD is not. The moderately conservative threshold of 2.5 times the MAD (Leys et al., 2013) was used to detect outliers over all critical stimuli (Median=546, MAD=166.8). This value included exactly all negative RTs – that is RTs where the button was pressed prior to target word presentation – as these are evidently wrong, 2.5x MAD also provides support for removing high RT outliers as well. Furthermore, in some sentences the target word was repeated later in the sentence, if a participant missed the first instance they may have reacted to the second one, which would also be captured as an outlier by 2.5 times the MAD. In total, 238 outliers were pruned leaving a final dataset of 1926 responses. Table 4.2 gives the mean RTs, standard deviation, number of stimuli for the different conditions and the difference in RT between FP and SP conditions.

A two-way ANOVA over the by-subject mean scores per condition showed a significant effect of pause type (F(1, 29)=12.73, p<0.005), no effect of speech type (F(2, 58)=0.805, p=0.452) and an interaction effect of pause and speech types (F(2, 58)=8.359, p<0.001). After applying Bonferroni correction it was found that the RT was significantly faster in both the natural and vocoded FP conditions than in the corresponding SP conditions (see Table 4.2), however no significant differences between the synthetic conditions existed. Exploring the interaction, it was found that in the SP conditions the RTs for synthetic speech were significantly faster than for vocoded speech (t(636)=3.778, p<0.005) and marginally faster than for natural speech (t(645)=2.729, p=0.059). Overall, the results of Fox Tree (1995, 2001); Fox Tree and Schrock (1999) have been replicated, and it has been shown that current vocoding techniques are able to represent the acoustic cues that are used by listeners to react more quickly to a target word after an FP. However, this effect is not replicated when hearing synthetic speech. In fact, in both synthetic FP and SP conditions, RTs are found that are similar to the RTs found in the other FP conditions, with a tendency toward the SP condition eliciting faster RTs – showing that there is currently no advantage to including FPs in synthetic speech.

	Mean RT (SD) in ms	Difference in ms	Adjusted p	N
Natural FP	532.4 (146.9)	-38.9	<0.05	312
Natural SP	571.2 (150.8)			312
Vocoded FP	541.5 (146.2)	-41.5	<0.005	322
Vocoded SP	583.0 (146.3)			303
Synthetic FP	554.5 (140.9)	14.6	= 1	342
Synthetic SP	539.9 (141.8)			335

Table 4.2: Mean reaction times (RT) with standard deviations (SD) in ms for filled pause (FP) and silent pause (SP) conditions for the three types of speech and the RT difference from FP to SP conditions. N is *after* outlier removal.

4.3.4 Discussion

This initial experiment replicated earlier findings which were based on natural speech, showing that participants are quicker to identify a target word when it follows a FP than when it follows a silent pause of equal length. This finding extends to vocoded speech in a straightforward manner, but not to synthetic speech. This raises the question: what is it about synthetic speech that inhibits the appearance of the effect? Our results show that vocoding is not the issue even though a slight decrease in general speech quality is to be expected (Hu et al., 2013). A similar results was found in Wester et al. (2015b) in which FPs, SPs and a tone was use for a temporal delay – vocoded speech replicated natural speech findings, but not synthetic speech.

One issue present in the creation of the synthetic speech, as compared to the natural and vocoded speech, was that of speaking rate (SR). While it was ensured that the durations of the FP and SP conditions within each speech type were the same, it was not ensured that the durations of the synthetic were matched to that of the vocoded and natural. Where the vocoder will produce the same durations as natural speech (within 5ms accuracy), the HMM synthesis method models duration based on the available training data. Notably the synthesis system which we used was trained on read prompts and it has been shown that spontaneous speech tends to have a faster speaking rate than read speech (Andersson et al., 2010b; Blaauw, 1992), as can be also seen in the corpus study in Section 4.2. Therefore, the synthetic speech durations were in general longer than the natural speech durations. Furthermore, one of the roles of FPs is to signal upcoming new information (Corley et al., 2007) and it can be argued that at lower SRs this role of FPs is superfluous as new information will be integrated fast enough

	Mean	SD	Difference (ms)	p	N
Natural	3.921	1.163	0.353	<0.05	77
Synthetic	3.569	0.581			75

Table 4.3: Mean speaking rate (in syllables per second) for natural and synthetic speech critical stimuli.

without the need for the additional marker of an FP.

To confirm this, a comparison of the SR of the natural and synthetic utterances was carried out (vocoded was left out as it would be equivalent to natural). SR was defined as syllables per second up to the target word. This definition ensures that spurious changes in SR after the critical point do not influence the overall SR and yields a measure of the time available to participants to react in. The SR of the natural speech was significantly higher than that of the synthetic, see Table 4.3. Furthermore, in a rough comparison of the data, the natural sentences were split into quartiles based on SR and it was found that the slowest quartile RTs were at least 30ms slower than their faster counterparts (Q1: 563ms, Q2: 512ms, Q3: 521ms, Q4: 532ms). Notably the SR of the synthetic speech falls within the lower quartile range. While there is not enough data to calculate reliable statistics, it suggests a trend in which it is possible that the FP advantage only appears at higher SRs not present in the synthetic speech. If that is the case, then this slower SR in synthetic speech may have affected results.

4.4 Speaking Rate, Reaction Time & Filled Pauses in Natural & Synthetic Speech

Due to the above difference in SR a second experiment was run, in which the overall SR of the synthetic speech was controlled to match the natural speech. The goals of this second experiment were i) to find whether speech rate could explain the lack of effect of FPs on RT in synthetic speech and ii) to investigate the effect of FPs in natural speech at different SRs, which has not previously been investigated. For simplicity, and because of its similar effect to natural speech, vocoded speech was not included in this experiment.

	FP	SP	Difference (ms)	Adjusted p
Natural	511.6	533.2	-21.5	< 0.05
Synthetic	547.7	524.5	23.2	< 0.05

Table 4.4: Mean RT for filled pause (FP) and silent pause (SP) conditions for natural and synthetic speech (SR combined) and the RT difference from FP to SP conditions.

4.4.1 Data

Selecting from the same AMI corpus as before, 80 critical and 40 filler stimuli were chosen which included either a repetition or 'UH'. The exclusion of 'OH' was due to the small number of sentences with 'OH' in the previous experiment, its relative low frequency in the corpus in general, and to simplify the experimental setup. The critical stimuli were chosen to represent three speaking rates, fast, medium and slow. We used the speech from the previous experiment as a guide, faster sentences were chosen to match faster natural speech (Fast), slower sentences to match the generally slower synthetic speech (Slow) and a medium category to match the average speaking rate in natural speech (Medium). The total duration, excluding initial and final silences, of each natural stimulus was measured and used to define the length of the synthetic stimuli. To avoid further editing of the synthetic speech, it was decided not to simply stretch or compress each synthetic stimulus to match the natural, but rather to require the stimuli to be a certain total duration at synthesis time. Again, this allows the system to deviate slightly from the prescribed overall sentence duration, but it ensures the system produces as natural an utterance as it is capable of. Note, however, that internally in the sentence the phone durations were not matched to that of the natural speech (i.e., natural phone durations were not used) – here only the total sentence duration was matched. All experimental materials are available at the thesis repository (Dall, 2017).

4.4.2 Method

Thirty-two native English speakers (mainly students at the University of Edinburgh) participated in the experiment. None of them had participated in the first experiment. The experimental procedure was similar to that in the previous experiment.

4.4.3 Results

Due to experimenter error three natural sentences were incorrect and removed from the analysis. Of the remaining 2496 critical observations, 406 were null responses and pruned. Using the 2.5 MAD threshold to detect outliers (Median=514, MAD=139.36), a further 205 responses were pruned. A two-way ANOVA over the by-subject mean scores per condition was run. For pause and speech type it showed no significant effect of pause type (F(1, 30)=0.098, p=0.757), a significant effect of speech type (F(1, 30)=5.112, $p<0.05$) and an interaction effect of pause and speech types (F(1, 30)=22.19, $p<0.001$). Investigating the speech type effect, after applying Bonferroni correction, we see that natural speech results in a mean faster RT of 13.7ms ($p<0.05$) compared to synthetic speech. Exploring the interaction effect (see Table 4.4) we find that FPs in natural speech, as in Experiment 1, result in *faster* RTs, however for synthetic speech they result in *slower* RTs. That is, we again have a benefit of FPs in natural speech, but we now see the opposite effect in synthetic speech with FPs resulting in slower RTs. Looking at the SR effect we find an overall effect of SR (F(2, 61)=6.083, $p<0.005$), an interaction of SR and speech type (F(2, 61)=3.770, $p<0.05$), no interaction with pause type (F(2, 61)=0.016, p=0.984) and no interaction between all conditions (F(2, 61)=1.656, p=0.199), see Table 4.5 for an overview. After applying Bonferroni correction the overall effect of SR is that we find slower RTs in the slow speed condition compared to the medium ($p<0.05$) and fast ($p<0.01$) but no difference between medium and fast (p=1). So, slower speech results in slower RTs for participants. For the interaction effect, we find that RTs in the synthetic conditions are generally slower than in the natural, except for in the fast condition. This is due to the natural fast SP condition (Figure 4.3) which is the only condition that does not follow the general pattern of RTs becoming faster as the speech becomes faster.

4.4.4 Discussion

Again, FPs in natural speech provide a benefit in terms of faster RT, however for synthetic speech the opposite effect now appears with FPs resulting in slower RT. Thus, the slower SR of synthetic speech compared to natural speech was not the reason for the lack of an effect in synthetic speech in the first experiment. Rather we see that a slower SR results in slower RTs in both natural and synthetic speech. This is unexpected: if RTs represent a measure of comprehension time we would expect slower speech to have at least as fast, if not faster, RTs than faster speech. It is possible that

	Mean SR (syll/s)	**Mean RT (ms)**	**RT SD (ms)**	**N**
Natural FP				
Slow	2.607	520.3	125.6	172
Medium	3.817	511.9	128.6	138
Fast	5.561	499.5	133.1	126
Natural SP				
Slow	2.650	537.4	108.7	173
Medium	3.846	523.2	124.7	130
Fast	5.551	537.9	123.4	119
Synthetic FP				
Slow	2.751	563.0	129.6	134
Medium	4.021	545.1	115.8	162
Fast	5.173	539.7	121.9	204
Synthetic SP				
Slow	2.702	550.3	122.2	150
Medium	4.007	522.4	121.1	191
Fast	5.098	505.8	122.6	186
All				
Slow	2.672	542.8	121.4	629
Medium	3.935	525.7	126.4	621
Fast	5.299	520.7	129.6	635

Table 4.5: Overview of speed divided conditions. SR = Speaking Rate in syllables per second.

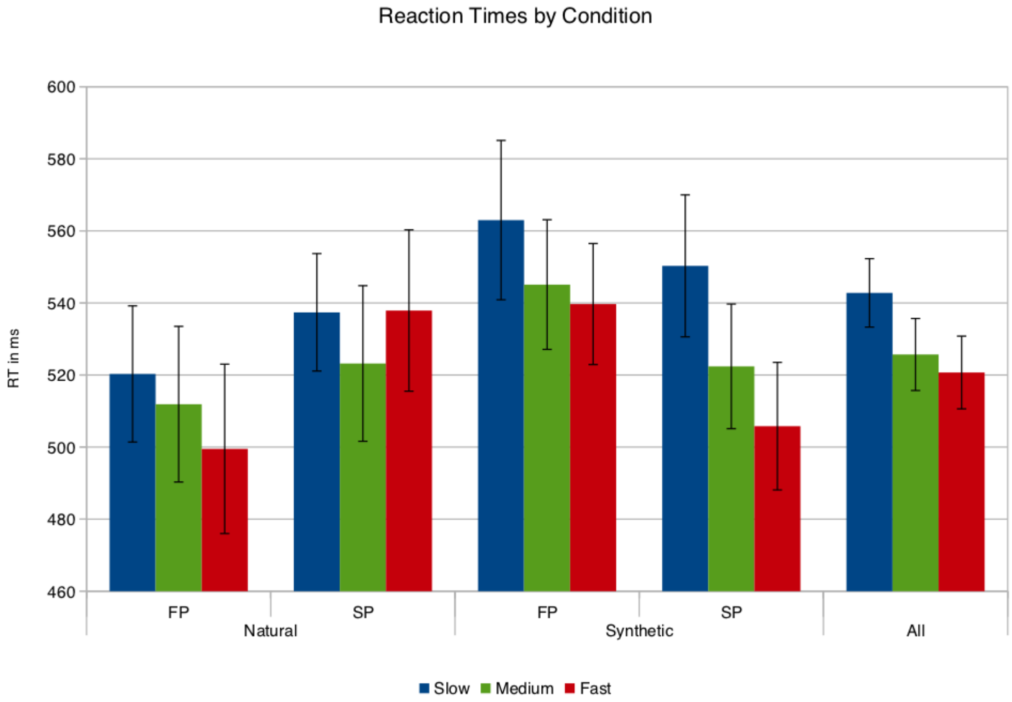


Figure 4.3: Mean RTs over each condition and speaking rate category in ms. Error bars show 95% confidence interval.

people adapt their processing speed to the rate of incoming information and thus the slower speech yields overall slower processing. While the results could be interpreted as showing faster speech to provide a benefit, this is probably only true around normal SRs, despite the Fast condition generally resulting in faster RTs than the Medium condition. Rather, it is likely that there is a ‘sweet’ spot SR range, around normal conversational SRs, in which we see the lowest RTs – when speaking much faster, listeners are likely faced with intelligibility issues which would hamper RTs and result in more null responses.

While SR did not account for the difference, it is possible that the nature of the natural SP condition did. Where the synthesis system creates the sentence to a given specification, for the natural speech the FPs were digitally edited out and replaced with a pause. This editing may have influenced our findings. To test this, each of the 80 critical sentences from the second experiment were used in a spot-the-edit test. Two groups of 8 participants were presented with 10 critical and 10 filler stimuli. In the first group, none of the stimuli had been edited, and in the second group the critical sentences were edited. The rate at which subjects believed an edit to be present did not differ (group with edits: 35%, group with no edits: 31%) and of guessed edits only 67% were correct, suggesting subjects were unable to correctly identify the edits. While it is possible that the edits may still have had a subconscious effect, it is beyond the scope of this paper to test this. Considering that similar testing of the methodology has been done Fox Tree (1995, 2001); Fox Tree and Schrock (1999), and splicing in FPs instead of SPs results in the same effect Fox Tree (2001), this seems unlikely to be the reason.

4.5 Overall Discussion and Conclusions of Reaction Time Experiments

FPs in synthetic speech do not behave in a similar manner to FPs in natural speech. Where natural FPs provide a benefit in terms of faster RT to a target word compared to a pause of equal length, synthetic FPs give rise to the opposite effect, namely a slower RT. This is not due to the effect of vocoding: vocoded speech follows the same pattern as natural speech. It was tested whether the generally slower speaking rate of synthetic speech caused the effect to appear and found this not to be the case. In fact, it was found that a slower SR tended to produce slower RTs *also* in natural speech; this is a new effect which has not been reported in the literature before. Another potential reason, other than SR, for the different results for the synthetic speech, is that the synthesised FPs are of a much lower quality than the surrounding speech, e.g. some are very long while most are very short.[6] This is likely due to the nature of the training data for the synthesis system. As mentioned earlier, current synthesis systems rely on recordings of read aloud scripted prompts which do not contain any FPs at all. While HMM synthesis is known for its greater robustness to missing training data, compared to concatenative synthesis (Jurafsky and Martin, 2008), a particular problem appears in the representation of FPs. FPs are realised quite differently acoustically than standard phones (as described in Section 4.2), but the training data for a standard voice contains no training samples of these different phones. As such, without training samples the system will attempt to infer the synthesis parameters from, e.g., mid-word and mid-syllable phones not like those found in an FP (see Chapter 7 for evidence of this). While repetitions are realised differently depending on the word that is repeated, similar acoustic differences to those found for FPs between words and their repetitions have been noted by Adell et al. (2006). One could attempt to mitigate this by including FPs

[6] Samples available at the thesis repository (Dall, 2017)

in the scripts for recording, however, as shown in the previous chapter, as spontaneous speech is preferred over read prompts it seems likely that these FPs would not be as well received as spontaneously produced FPs. The proposal is therefore to follow Andersson et al. (2012) in training HMM-voices from spontaneous speech which includes natural examples of FPs (see Chapter 7). Potentially this would not only provide the necessary data, but also naturally speed up the synthetic speech due to the generally higher SR in spontaneous speech, removing the need to enforce specific duration requirements on the synthesis system. See Chapter 10 for a revisit of this experiment in the context of improved FP synthesis.

To conclude, FPs result in faster RTs in natural and vocoded speech, but slower RTs in synthetic speech. SR did not account for this difference, however a tendency for RTs to slow down in response to slower speech was present, slower SR being the norm in synthetic speech. To enable speech synthesisers to show the same effect, the proposal is to include FPs in the training data, which should produce both better FPs and an increased SR.

4.6 Filled Pauses in Change Detection in Natural, Vocoded and Synthetic Speech

Following on from the previous two experiments, this experiment explores the effect of FPs in a change detection paradigm. While the reaction time experiments provide evidence that FPs affect peoples' on-line processing, FPs may have other, and longer term, effects. Change Detection (Sturt et al., 2004) is a paradigm in which participants are asked to listen to short paragraphs of speech and are subsequently presented with the contents of the speech in writing. It is then the task of the participant to detect if a single change has occurred in the text, compared to the speech. This requires participants to not only process the speech as it is heard, but also to memorise it long enough to detect a change at a later point. Thus change detection, as opposed to reaction time, experiments provide a measure of the "memorability" of the speech over a slightly longer term.

The basic effect reported by Collard (2009), Chapters 6 & 7, is that the presence of an FP prior to the changing word, as compared to fluent speech, increases the change detection rate by 10-15%. Collard (2009) concludes that the acoustic quality of the FP is responsible for the effect (Chapter 7.6, pp. 128). His conclusion was based on manipulating silences around the FP but (Sanford and Molle, 2006) has shown that a simple silent pause can make the same effect appear, something also found in other related studies (Corley and Hartsuiker, 2011). Therefore this investigation extends previous work by including silent pauses and a discourse marker ('like') in addition to the FP 'UH' in natural speech – to see if the effect is unique to FPs. Furthermore, as the primary interest is in the effects of FPs on listeners in relation to synthetic speech, the experiment is performed using vocoded and synthetic speech in addition to natural speech.

The working hypothesis is that a similar pattern to the RT experiments would appear, in which the effect of disfluencies is present in natural and vocoded speech, but not in synthetic. This is motivated by the results of the prior experiments, but also by the assumption that current vocoding techniques do not degrade the quality of the speech in a way which would prevent the effect from appearing. It is possible however, that a differing pattern will appear due to the differences between the two paradigms. In RT experiments we are testing people's online monitoring and recognition of speech, whereas in change detection people are required to memorise the heard speech in order to detect the change at a later point. Even though participants may understand the speech, they may not be able to effectively memorise it.

4.6.1 Data

To perform the change detection experiment, 35 short paragraphs, 16 critical, 16 filler and 3 practice, said by the same speaker in a spontaneous conversation were prepared – these were not from the speaker recorded in the previous chapter, again due to these experiments taking place prior to the corpus being ready for use. Instead these were provided by Martin Corley, having been used for previous change detection experiments.

In each paragraph, a target word was chosen and four alternative paragraphs were created. One where the target was preceded by a FP ('UH'), a silent pause (SP), the discourse marker 'like' (DM) or by nothing (i.e. fluent speech). The original paragraph was of one of these four cases, and the alternatives were made by altering the original by splicing out the segment immediately preceding the target word and splicing in the relevant replacement. This was done in a similar manner to the previous two experiments in this chapter and, as these were found not to be noticeable, no test of this was performed on this data. The change word was a near-synonym or semantically related

Sample Paragraphs
Last week I was thinking actually about writing a book. The possibilities are endless, but I came down to it and thought, well, I'll start with a *cemetery/graveyard* setting, and, within a few minutes, the whole book started to write itself.
My parents are quite possibly the biggest alcoholics ever. And when I went home a couple of days ago they kicked off the evening with a smooth cocktail, or so they called it, which is a drink comprising of *lime/lemon*, and orange juice, vodka and a splash of Malibu.
She was so scared of spiders I had to take the almost parental role, and every time I heard the scream I knew to take a glass and run in to try and find her. We had this kind of *agreement/arrangement* that she would do certain things around the kitchen and I would save her whenever there was a creepy crawly.

Table 4.6: Sample paragraphs presented to participants. The first of the italicised words is what is heard and the second what is seen by the participant.

to the target word (i.e. the close-change condition of Collard (2009)). For the filler stimuli, no change existed, however a dummy target word was still chosen in front of which either an FP, SP or DM was placed. This was done to ensure that participants did not accustom to the FP, SP or DM being a marker of changes to happen. Furthermore, the paragraphs potentially included other FPs, DMs and SPs than the critical one so participants could not learn to use those as cues for the change. Of the practice sentences, two contained no change and one a change. Table 4.6 illustrates a few paragraphs. All experimental materials available at the thesis repository (Dall, 2017).

The vocoded versions were created by taking the natural paragraphs and vocoding them using STRAIGHT (Kawahara et al., 1999b); no further modification to the audio was made. The synthetic utterances were made using HTS 2 (Zen et al., 2007a) and a good-quality state-of-the-art HMM-based British English Female voice trained on approximately 8 hours of speech – the same as that used in the previous RT experiments. The transcripts of the paragraphs were used for the synthesis, and versions including a FP or DM was made by inserting these as words in the token stream, whereas the SP version was made in a similar way as in the RT experiments, the length of the SP was thus similar to that of the FP.

4.6.2 Method

108 participants were recruited: 36 listened to natural speech only, 36 to vocoded and 36 synthetic speech. Each participant only heard samples with either an FP, SP or DM such that for each type of speech and each type of pause there were 12 participants. Each participant listened to the practice sentences and then to each of the 32 paragraphs in a random order, of the 16 critical half contained the appropriate form of pause, and the other half no pause (with 6 participants getting one set and other 6 the other set). In total this yielded 576 (36*16) critical evaluations per speech type and 192 (12*16) per condition (FP, SP or DM) within each speech style. Samples and experimental materials are available at the thesis repository (Dall, 2017).

4.6.3 Results

Due to an error in the experiment scripts, 96 trials were invalid (5.5%) and were removed from the analysis. In 116 of the remaining trials (7.1%) participants correctly detected a change but incorrectly specified which change. In 16 of these the participant answered that the DM was the change which can arguably be considered correct as it was not presented in the text. Therefore, two analyses were carried out - with (Exact) or without (Permissive) the exact specification of change. Notably, the patterns of the results are identical. Please note that in the following analysis *disfluent* speech includes FPs, DMs and notably SPs, *fluent* speech is thus only speech with none of these present. This was done as that is the original experimental methodology (Collard, 2009), however, in Section 4.6.4 each “disfluent” speech type is discussed separately.

A two-way ANOVA over the by-subject mean scores per condition was run. There was no overall effect of disfluency type (FP, DM, SP) or disfluency condition (Fluent or Disfluent), however a significant effect of speech type (Permissive: F(2, 99)=5.917, $p<0.005$, Exact: F(2, 99)=10.377, $p<0.0001$) was found. An interaction effect was also found between speech type and disfluency condition for the exact analysis (F(2, 99)=5.180, $p<0.01$) which was only marginal in the permissive (F(2, 99)=2.788, p=0.066). After applying Bonferroni correction the effect of speech type is such that, for the natural speech, detection rates were significantly *higher* than vocoded (Permissive: t(139)=2.692, $p<0.05$, Exact: t(140)=4.745, $p<0.0001$) and synthetic (Permissive: t(142)=3.878, $p<0.001$, Exact: t(139)=4.699, $p<0.0001$), but no difference existed between synthetic and vocoded speech (Permissive: t(138)=0.870, p=1, Exact: t(133)=0.662,

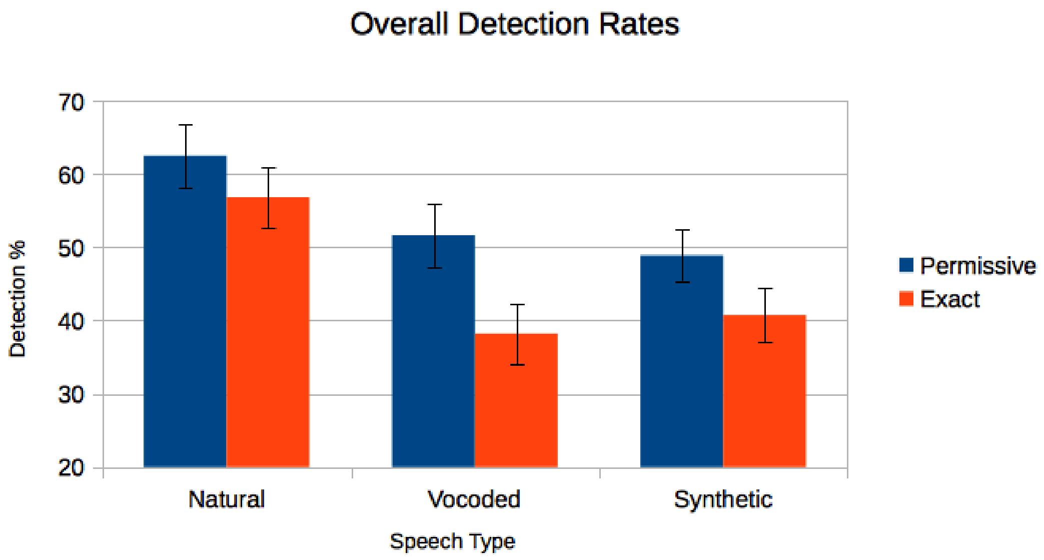


Figure 4.4: Detection rates per speech type. Permissive includes correct detection of change but incorrect identification. Exact does not.

p=1), see Figure 4.4. That is, changes are generally detected better in natural speech than in synthetic (by 13.6% in the permissive and 16.1% in the exact case) and vocoded (by 10.9% in the permissive and 18.6% in the exact case).

The interaction effect (see Figure 4.5) was explored as it was significant in the Exact case and near significant in the Permissive. After applying Bonferroni correction, there was no effect of disfluency condition in synthetic (Permissive: t(70)=1.374, p=0.521, Exact: t(70)=0.582, p=1) and vocoded speech (Permissive: t(70)=0.355, p=1, Exact: t(70)=0.075, p=1), however a significant effect was present in natural speech (Permissive: t(70)=3.326, p<0.005, Exact: t(70)=3.307, p<0.005). The presence of a disfluency did not have any effect on detection rates in synthetic and vocoded speech, however in natural they increased detection rates by 14.4% in the Permissive and 15.3% in the Exact case.

As disfluency had an effect in natural speech, individual tests for each disfluency type were run compared to the no-pause condition. After applying Bonferroni correction, a marginal effect of the FP was found in the permissive case (t(220)=2.356, p=0.058) which was significant in the exact (t(220)=2.468, p=0.043). For the DM, a significant effect was found in the permissive case (t(223)=2.736, p=0.020) which was marginal in the exact (t(223)=2.3608, p=0.057). There was no effect of SP in the Permissive case (t(236)=1.739, p=0.250) but a marginal effect in the Exact (t(236)=2.234,

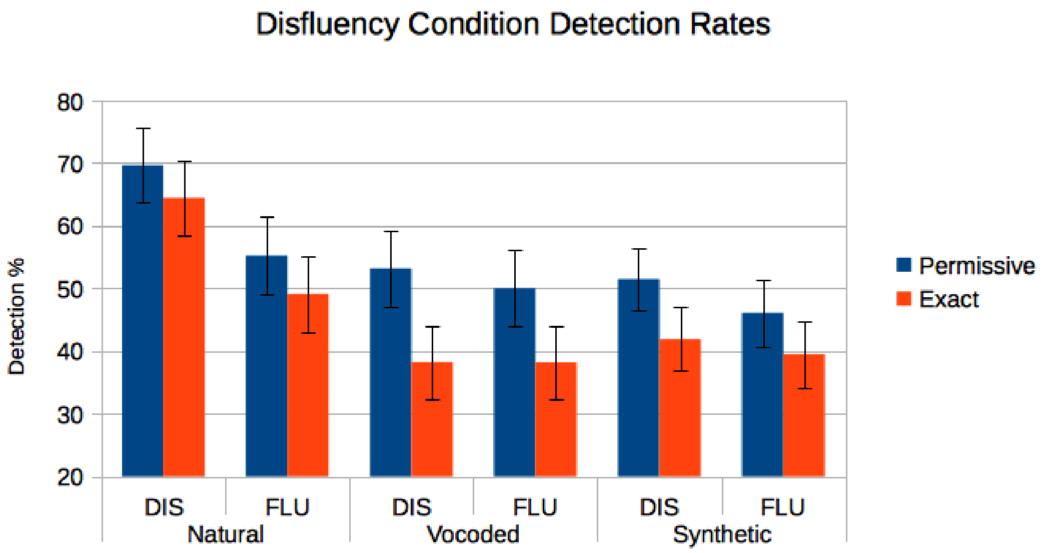


Figure 4.5: Detection rates divided by disfluency condition and speech type. DIS are disfluent conditions and FLU the fluent condition.

p=0.079). Thus, both the FP and DM significantly increase detection rates, whereas the SP seemingly has a less clear-cut effect. See below for a discussion about this. Figures 4.7, 4.8 and 4.6 show individual detection rates for each disfluency type in each speech type.

4.6.4 Discussion of Change Detection Experiment

Disfluent speech increases change detection rates in natural speech compared to fluent speech without disruptions. However, this is not the case for vocoded or synthetic speech (Figure 4.5).

In natural speech, the FP and DM provide the larger and more significant benefit, whereas the contribution of a SP is less clear-cut (Figure 4.6). The results are, seemingly, in line with Collard (2009) who concludes that the acoustic quality of the FP is important in providing a benefit. While Collard investigated varying the length of SPs surrounding the FP he did not evaluate SPs on their own, as done here. The SP results here are, however, different from those found by Sanford and Molle (2006) in a very similar experimental setting. Their results are in line with the temporal delay hypothesis of Corley and Hartsuiker (2011) that it is simply the disruption which causes the increase in change detection rates. This is something which, in a strict interpretation, is not supported by these results. While the tendency was for the SP to have lower detection rates than either FP or DM it did still increase detection rates (Permissive: 9%, Exact: 15%). It may be that the difference between the FP/DM and SP results is a consequence of there being many comparisons and as such statistical power has been lost. That the effect appears with the DM can support both the hypothesis that it is the disruption which is the cause but also the idea that the use and purpose of DMs and FPs is similar (e.g. as seen in Fox Tree and Schrock (1999)). To determine which is more likely to be true using a non-speech condition as in Corley and Hartsuiker (2011) could be considered in future studies, while also replicating the SP experiment with a focus purely on natural speech could help. These experiments are, however, outside the scope of this thesis.

Current synthesis and vocoding techniques do not produce speech for which the change detection results observed for natural speech are replicated (Figure 4.8 and 4.7). Where FPs, DMs and SPs increase the detection rate by 11-17% in natural speech there is no discernible pattern in synthetic and vocoded speech, rather, they tend to produce the same detection rates. Not only did the natural effect not appear, for both vocoded and synthetic speech the overall detection rate dropped as compared to natural speech by 11 to 18%. This is not just an effect of increased detections in the disfluency conditions of the natural speech, but rather an overall effect of the speech type. It is notable that this inability to replicate the effect occurs in *both* synthetic and vocoded, as the initial expectation was that current vocoding techniques were good enough to replicate the effect – as in the RT experiments earlier. That they do not, suggests that it is not simply a matter of the speech prosody and general naturalness being poor, but rather that there is something about the inherent speech quality of the vocoder which limits synthetic speech in this regard.

In the reaction time experiments, the vocoded speech elicits the same patterns as natural speech, which is in contrast to current results. Vocoding is known to introduce a buzzy character to the speech, while we are aware of the effects on perceived naturalness of this (Henter et al., 2014), other possible psychological effects of this buzziness are unknown. It is possible that we have found one of them. To detect a change, the participant must necessarily be able to commit to (short term) memory what was being said in the paragraph in order to compare with the text later. Thus if the effect of vocoding decreases participants ability to memorize the salient elements of the paragraph, it should show an overall decrease in a participant's ability to detect changes, something which is the case. This decrease is likely due to an additional strain on the participant's cognitive resources and can also explain the lack of disruption/temporal

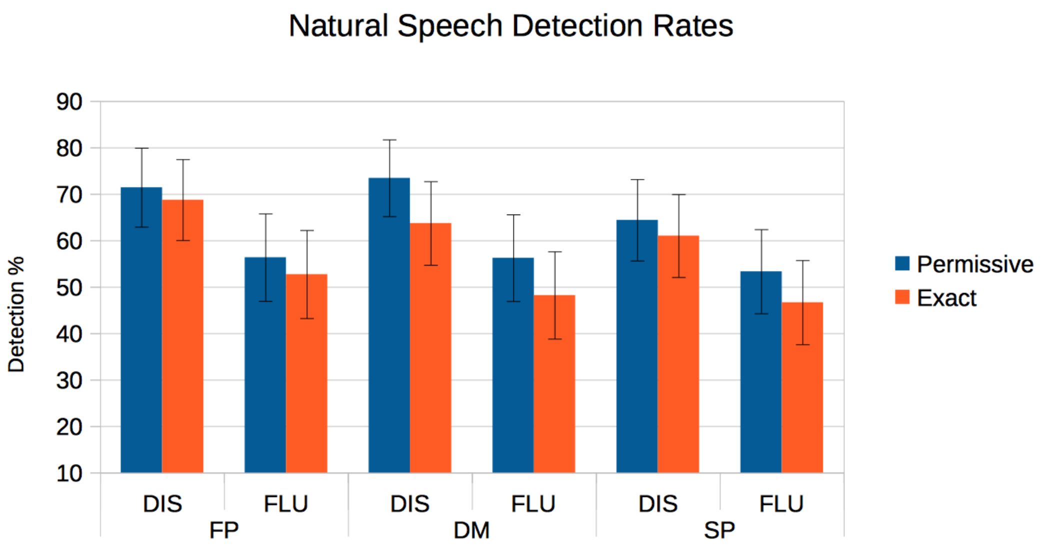


Figure 4.6: Detection rates per disfluency type for natural speech. DIS = disfluent. FLU = fluent. FP = filled pause. DM = discourse marker. SP = silent pause.

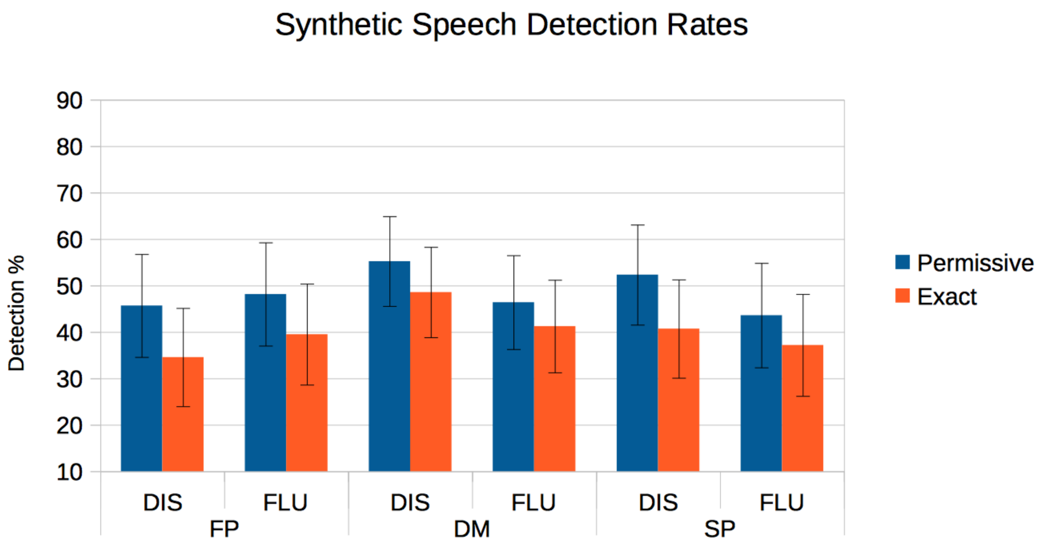


Figure 4.7: Detection rates per disfluency type for synthetic speech. DIS = disfluent. FLU = fluent. FP = filled pause. DM = discourse marker. SP = silent pause. FLU = fluent. DIS = disfluent.

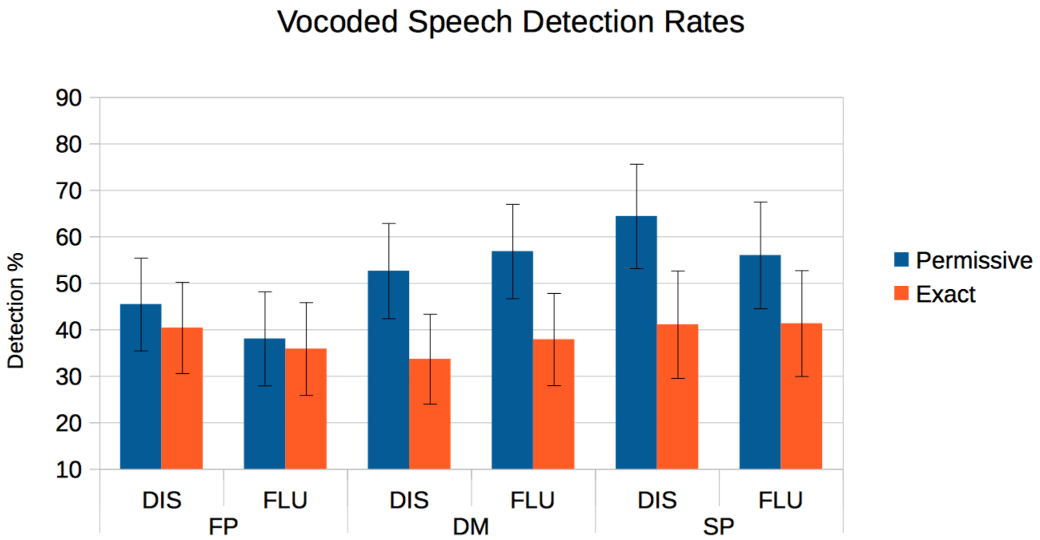


Figure 4.8: Detection rates per disfluency type for vocoded speech. DIS = disfluent. FLU = fluent. FP = filled pause. DM = discourse marker. SP = silent pause.

delay effect. The participant must use so many resources to simply process the incoming speech stream that any potential benefit to be had from the disruption is lost. Following Collard (2009), the effect of disfluency found in natural speech is due to heightened attention to the target word, resulting in better recall and notice of changes. While durational and prosodic cues may still be present after vocoding, if the participant is already straining their cognitive resources to simply understand and commit the content to memory, it is likely that these cues do not result in an attentional shift. This is, however, speculative and further experimental evidence would be needed. Experiments explicitly manipulating the cognitive strain on participants, such as dual-attention tasks, could be used in combination with a change detection paradigm using natural speech, if this alters the results for natural speech to look similar to those of vocoded and synthetic speech it would provide evidence for a cognitive strain hypothesis. These experiments are, however, outside the scope of this thesis as the focus on synthetic speech would be lost.

4.7 Chapter Conclusions

In this chapter, investigations into the potential subconscious benefits of including filled pauses in TTS were described. A brief review of corpus studies, and a presentation of a short new study, show the ubiquity of FPs in spontaneous speech and regularities in their use and realisation. Following this, a reaction time experiment investigating the effect of ‘UH’, ‘OH’ and repetitions was presented using natural, vocoded and synthetic speech. The effect, faster reaction time to a target word in the presence of an FP as compared to a silent pause, was replicated in natural speech, and straightforwardly appeared in vocoded speech. However, for synthetic speech the effect was not present. A follow-up study controlling for speaking rate was done in which it was found that there was a general effect of higher SR resulting in faster RTs. However, this effect did not mask any effect of FPs in the synthetic speech. In fact, the second experiment highlighted that FPs, in current synthetic speech, result in the opposite effect – slower RTs. A similar investigation into the effect of vocoding and synthesis on FP benefits in a change detection paradigm, showed that, in this paradigm, vocoding was a limiting factor as the effect was found in neither vocoded nor synthetic speech. This result is in contrast to the results of the RT experiments presented here and those of Wester et al. (2015b) in which the vocoded speech also replicated natural results. The suggestion is that the two tasks (Wester et al. (2015b) also used an RT paradigm) differ in what is asked of participants – RT experiments deal with online processing whereas change detection also relies on memory – and that this reliance on memory may have been disrupted due to the effects of vocoding.

A potential study to determine if this cognitive strain hypothesis is true has been outlined, but it will not be performed in this thesis. That the vocoded speech exhibit the same effect as natural speech in the RT experiments show that this effect should be replicable using synthetic speech, and in Chapter 10 the experiment is re-run using the improved FP synthesis methods proposed in this thesis. Going forwards, in terms of FPs, the focus will thus be primarily on improving FP realisation because this is an issue when looking at the details of the FPs produced by the standard system employed in these tests.

As discussed in Chapter 3 current techniques do not produce good synthesis when purely using spontaneous speech, however, as shown in this chapter – state-of-the-art read speech-based voices cannot replicate psycholinguistic findings at least partly due to bad FP synthesis and potentially due to bad prosody. Therefore, we will now turn our attention away from FP realisation (which we will revisit in Chapter 7) and psycholinguistics (which we will revisit in Chapter 10) and focus on improving both spontaneous and read speech modelling. This focus is clearly needed as better modelling must be done before we can replicate psycholinguistic findings.

Chapter 5

Pronunciation Variant Forced Alignment

A Note About Collaborative Work

The work presented in this chapter was done in collaboration with a number of people. Sandrine Brognaux (University of Mons and Universite Catholique de Louvain) was the second transcriber of the Gold standard manual transcriptions, helped define the manual set of pronunciation variation and participated in general discussion. Korin Richmond (CSTR) implemented the initial system for producing the lattices containing the pronunciation variants and participated in general discussion, particularly regarding which manual rules to implement. Junichi Yamagishi (CSTR) created and provided the large female average voice model used for model initialisation of the Gold standard system. Cassia Valentini-Botinhao (CSTR) was the third transcriber, and Rob Clark (CSTR/Google) provided help and advice regarding the multisyn build tools. Julia Hirschberg (University of Columbia) participated as a discussion partner for Sandrine Brognaux while she was doing the manual transcriptions and rules derivation. Parts of this work have been published as Dall et al. (2016a) but it is significantly expanded on here, particularly in terms of finding the best alignment method.

5.1 Introduction

So far, I have reported on experiments showing the feasibility of utilising spontaneous speech data for TTS and its potential naturalness gains (Chapter 3) and the potential, subconscious benefits of using disfluencies in speech (Chapter 4). It has been shown that a read speech-based voice is preferred over a spontaneous speech-based voice and also demonstrated the inability of current, read speech, TTS systems to replicate the subconscious effects of disfluencies in natural speech (Chapter 4).

In this chapter, I will present a method for utilising pronunciation variation to better annotate the training corpus. This is motivated by the observation in Chapter 3 that spontaneous speech-based voices suffer from bad individual phone modelling and also the observation in Chapter 4 that current read speech-based voices cannot replicate psycholinguistic findings. Pronunciation variant forced alignment has the potential to improve both issues as the variation in the speech, particularly reduced pronunciations in the spontaneous speech, can be better captured by allowing for these variants in modelling, which also has the potential to improve synthesis of phenomena such as FPs. Furthermore, an analysis of the initial accuracy of the forced alignment procedure on both read and spontaneous speech is presented, showing that the standard method produces many, and particularly for the spontaneous speech, serious errors (see Section 5.2). Correcting these errors is likely to improve the modelling of both read and spontaneous speech and should be desirable.

Forced alignment is an essential step to prepare the speech data for text-to-speech synthesis (TTS). It is the segmentation of raw speech waveforms into the phones of the utterance for use as units in unit selection synthesis, or to train models for statistical parametric speech synthesis (SPSS). Alignment is normally performed using this automatic method as opposed to manually aligning the speech, primarily because manual alignment is both expensive and error-prone (van Bael, 2007).

In English speech synthesis, the standard forced alignment procedure lets the TTS front-end produce a transcription which the algorithm is then forced, hence the name, to find boundaries of in the acoustics. This transcription may be incorrect and the phones wrong, e.g. when reductions or deletions occur. As such, the phones may not exist in the utterance, although due to the forced nature of the method these will still be “found”. This is not a major issue in unit selection as the join cost will discourage any badly aligned units from being selected. In SPSS a join cost is not used, but it is usually assumed that the transcription is sufficiently close to correct that this is not an issue. As a consequence, it is assumed that any bad units are either averaged out as “noise”, or that by being consistent across training and synthesis, errors will not affect the output speech negatively. This is what I will here call the consistency assumption between training and synthesis, namely that making the same mistakes consistently may “accidentally” have positive effects, such as appropriate phone reductions (Campbell, 1997).

To the best of my knowledge, the truth of this has not been tested in synthesis, and it is possible that this has been an extrapolation from automatic speech recognition in which manual alignments do not improve word error rates over forced alignment (van Bael et al., 2007). It may also be derived from experience in unit selection where, as said, the join cost will discourage badly aligned units from being used[1]. Kim (2004) provides evidence that the assumption holds in a unit selection paradigm, but does not discuss the finding as their focus was on retaining the consistency. Furthermore, for SPSS based on spontaneous conversational speech data, it is worth remembering that there are significant differences between the appearance of conversational phenomena in standard read prompts and spontaneously produced speech (Shriberg, 1996; Bortfeld et al., 2001; Fox Tree, 1995; Goddijn and Binnenpoorte, 2003; Brognaux and Drugman, 2014; Brognaux et al., 2014b).

Earlier work by Andersson (2013), Chapter 3, on spontaneous TTS admitted problems with speech alignment. This was solved through data selection, artificial stretching of the spontaneous speech and a proprietary alignment system of which we do not have the details. Unfortunately, no evaluation of the effect of this was performed, although from experiences in this thesis it may have been an important step.[2]

Also, recent studies in French by Brognaux et al. (2014a,b) have, again, demonstrated that there are significant differences between the occurence of conversational phenomena in standard read prompts and spontaneously produced speech, and, crucially, that correcting this can lead to improved synthesis quality. Using a corpus of sports commentaries (from Brognaux et al. (2013)) with hand-corrected alignments, an improvement in synthesis quality was achieved when using these manually corrected transcriptions for training and synthesis (Brognaux et al., 2014b). It is also worth noting that pronunciation variant forced alignment has long been used in automatic speech recognition and by researchers interested in speech segmentation, e.g., Binnenpoorte et al. (2004); Kessens et al. (1999); Paulo and Oliveira (2005).

These studies by Brognaux and colleagues show that manually corrected transcriptions can potentially benefit synthesis. It is however unclear whether this is due to the better phone accuracy in the alignment or due to a more natural pronunciation during

[1] Although if several badly aligned units occur in a row, issues may appear.

[2] On a side note: Artificially stretching the durations of the spontaneous speech is also a curious decision, it probably alleviates the alignment issue to some extend as each phone will receive more duration and thus the models will be less squashed, but it will likely make the output speech less conversational which seems to defeat the point.

synthesis. This chapters' focus is on the first part, while Chapter 6 will focus on the second.

While manually correcting automatically produced alignments is faster, and thus cheaper, than full manual transcriptions (van Bael, 2007) it is still far more time consuming and expensive than automatic alignment. However, if manual correction leads to better synthesis (Brognaux et al., 2014b) it is desirable that we can produce automatic alignments which are as good as manual correction, such that this manual correction can be avoided. Thus a method allowing for pronunciation variation at the training stage is developed and evaluated using synthesis based on the standard automatic transcription, i.e., breaking the consistency assumption between training and synthesis.

5.2 Forced Alignment Accuracy on Read and Spontaneous Speech

In order to test the hypothesis that spontaneous speech conforms less to standard transcriptions than read speech, the small corpus of 50 parallel read and spontaneous sentences from Chapter 3 was analysed. As a reminder, the corpus contains 50 sentences which were spoken in a normal conversation by a female British English speaker. These sentences were orthographically transcribed and given to the voice talent to read aloud as standard prompts (the talent was unaware that she had earlier said these sentences) to obtain read versions of them. The sentences thus contain exactly the same content and only vary in their acoustic realisation (read or spontaneous).

5.2.1 Gold Standard Alignment

To create a gold standard transcription, an automatic alignment of the two sets of 50 sentences was first obtained and then manually corrected. The automatic alignment was done using a large British English female average voice model created using the Voice Cloning Toolkit (VCTK) (Yamagishi et al., 2012) and adapting it to the read and spontaneous speech corpus from Chapter 3 respectively. The standard transcription was obtained using Festival 2.4 (Black et al., 2014) and the RP British English version of the Combilex dictionary (Richmond et al., 2009, 2010). For the spontaneous speech, the transcription also included pausing, information which was provided for the model as it helped alleviate the cascading issue (see Section 5.2.3). Alignments of the 50 parallel sentences were then obtained using their respective models.

These alignments were then, independently, manually corrected by two annotators[3]. Where these hand-corrections disagreed, the labellers met to discuss and agree upon a final transcription. If the disagreement was whether to keep the original Festival transcription or change it, the Festival transcription was often preferred. As noted by van Bael (2007) human labellers correcting automatic transcriptions are biased toward the initial transcription. Therefore, this agreed upon transcription is doubly biased *toward* the standard Festival transcription. In other words, it is a conservative measurement of the deviation from the actual realisation. In Section 5.4.4, the effect of this potential bias is investigated by having a third annotator correct the output of the variant pronunciation system proposed in this chapter. Furthermore, it is worth noting that the focus was on phone identity and not phone boundary; thus phone boundaries were only corrected if grossly incorrect, e.g. such as when a phone was deleted.

5.2.2 Transcription Accuracy Comparison

To evaluate the phone accuracies and inter-annotator agreement, the mean percentage deviation in Levenshtein distance (here also Phone Error Rate, or PER) was used with the manually corrected alignments as the gold standard (Table 5.1). While not the suggested method of van Bael et al. (2006); van Bael (2007) it is standard (Section 1.32, van Bael (2007)). This was used as a measure of transcription accuracy since it is quick and easy to determine, and the agreed transcription of the two original annotators constituted the gold standard for development. The Levenshtein distance is given by the minimal number of deletions, insertions and substitutions to transform one string into the other. So if comparing the string “k a t z” and “g r a t” the Levenshtein distance would be 3, one substitution (“g” for “k”), one insertion (“r”) and one deletion (“z”). Thus the Levenshtein distance (*LD*) between two strings (*a*, *b*) is:

$$LD(a,b) = min(\sum(insertions, additions, substitutions)) \tag{5.1}$$

Adapted scripts from the method in Chapter 9 of Brognaux (2015) was used to obtain the alignment between two phone strings and to calculate the distance based on this alignment. The adaptation consisted of not discriminating between any errors between phones, in Brognaux (2015) some phone substitutions were not penalized and neither was insertions and deletions of silences (p. 207 in Brognaux). Although

[3] Sandrine Brognaux and myself.

	Deletions	Insertions	Substitutions	Total	PER
Read					
Automatic Avg Voice	4	141	149	294	18.1%
Annotator 1	30	33	69	132	8.1%
Annotator 2	9	3	36	48	3.0%
Spontaneous					
Automatic Avg Voice	14	187	175	376	23.7%
Annotator 1	15	11	42	68	4.3%
Annotator 2	15	4	18	37	2.3%

Table 5.1: Overall differences between the agreed gold transcription, the standard automatic system and the annotators, with the gold transcription used as reference in each case. Avg = average. PER = Phone Error Rate.

arguably some phoneme substitutions may seem less harmful than others, such as devoicing compared to vowel/consonant confusion, in this work a very restricted set of possible changes are applied through the automatic methods and due to the transcribers bias toward the initial transcription (see Section 5.4.4) their transcriptions have few of this type of confusion.[4] Furthermore, as can be seen in Section 5.4.1, the correct detection of silences improves the resulting accuracy and so focusing on silence errors is important in this context. Thus, it seemed more appropriate to utilise the standard metric.

Table 5.1 shows the differences between the annotators and the standard phonetisation compared to the gold standard. The inter-annotator disagreement was 11.1% (read) and 6.6% (spontaneous) respectively. The automatic alignment is surprisingly bad, getting over 18% of all phones wrong for the “simple” case of clear read speech prompts and over 24% for the spontaneous speech. This is much higher than the inter-annotator disagreement. For the spontaneous speech, the automatic alignment is about 6% worse than for the read speech, and interestingly, the annotators are in much closer agreement for the spontaneous than the read speech. This is likely due to the spontaneous speech being further from the original alignment and thus when a change was made it was more likely to be agreed by both annotators that it was different from the original transcription, leading to it being less biased toward the original transcription.

The Festival transcription deletes very few phones, this is unsurprising because

[4] Note that Section 5.4.4 reveals a serious issue with this and Section 5.5 describes a perceptual evaluation, a more suitable test, of the resulting synthesis systems.

the standard transcription, which is very much based on the lexicon used, is generally the full pronunciation so it is to be expected that the alignment would not miss many phones actually present in the speech. It does, however, make a substantial amount of insertions and substitutions, particularly for the spontaneous speech.

Of the insertions, most are end of word stops, particularly “t” but often “d”, and the main substitutions are “t”s for glottal stops and end of word “z” for “s”. That is, assuming the annotator’s gold standard is closer to the actual realisation, we find that comparing that to the Festival transcription, deletion of end of word stops, glottalisation and devoicing is common. Together these account for 35% of all “mistakes”, or disagreements, with the annotators gold standard in the spontaneous speech and the pattern is similar for the read speech. What is notable here is that only the glottalisation could be considered speaker specific and non-standard in RP English (though common in many dialects). On the other hand, deletion of end of word stops and devoicing are generally common phenomena. In fact the main differences, between the annotator’s gold standard and the Festival transcription found here are similar to those in Fackrell et al. (2003).

5.2.3 Cascading Deletion Errors

Due to a much greater number of reductions and deletions in spontaneous speech, compared to read speech, some utterances experience an issue of cascading alignment errors illustrated in Figure 5.1. This is not an issue of poor boundary alignment, but of poor automatic transcription. The excerpt in Figure 5.1 shows part of the two words “basically because” in the read and spontaneous rendition. In the read speech, the automatic transcription is appropriate, but in the spontaneous the produced pronunciation is “basly ’cause”. This produces not only the problem of non-existent phones being “found” by the alignment procedure, but also the much more serious issue of phones being “pushed” later in the utterance, putting every single phone further down the line out of alignment. This creates phone examples for model training which are grossly wrong.

While the problem with because/’cause is arguably a lexicalised difference which could be resolved during an orthographic transcription, the issue of basically/basly is not, and such situations will cause problems in the trained models. For the gold standard phone transcription, lexicalised differences were considered orthographic transcription errors and corrected prior to automatic alignment. Furthermore, it was found

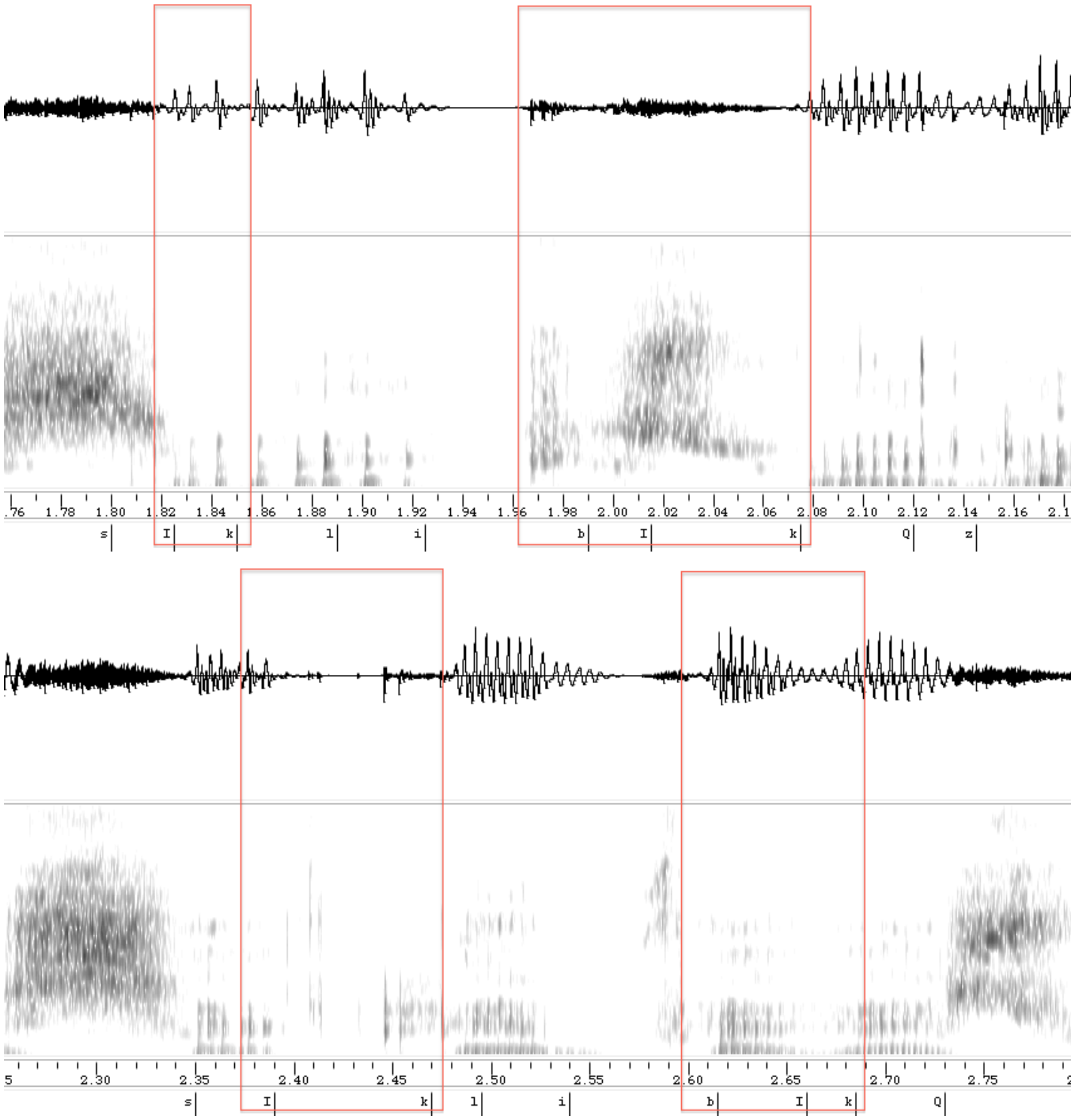


Figure 5.1: Segment of the same sentence produced spontaneously (top) or as a read prompt (bottom), showing issues with alignment in spontaneous speech. The red squares highlight the phones /I k/ and /b I k/ segment in both as found by the alignment. Notice how the /k/ in /I k/ is not present in the spontaneous speech but is in the read. Similarly, /b I/ are not present but rather just /k/ in /b I k/, this illustrates the issue with non-detected speech deletions in standard forced alignment.

that providing the transcribed pausing to the Festival system used as output for the manual transcription had the effect of limiting the cascading issue to only the next pause in the utterance. This was therefore done for the gold standard automatic alignment to ease transcription. This information was not, however, provided to the evaluated automatic systems below. In fact in Section 5.4.1 an additional silence removal step is added due to this difference.

5.3 Pronunciation Variation in Forced Alignment

To improve the phone identity accuracy on both read and spontaneous speech, a lattice-based forced alignment system was implemented based on the Festival multisyn voice building tools (Clark et al., 2004) and the underlying context-dependent rewrite rules of Combilex described in Richmond et al. (2007).[5]

As noted above, pronunciation variant forced alignment has long been used in automatic speech recognition and speech segmentation, and these methods are generally lattice-based. A lattice is often avoided in synthesis due to the consistency assumption, because the transcription found at training time cannot necessarily be produced during synthesis. The consistency assumption, however, seems to rely on the standard (Festival) transcription making only minor errors and as such is goo denough. This likely holds for standard read prompts, but, as described in the previous sections, not for spontaneously produced speech.

5.3.1 The Alignment Procedures

The standard forced alignment procedure used in this chapter proceeds as follows:

- Step 1: Initialise all HMM monophone models as 5 state models.
- Step 2: Estimate model parameters based on a flat-start.
- Step 3: Re-align based on obtained models and re-estimate HMM models.
- Step 4: Add short pause model, re-align and re-estimate.
- Step 5: Re-align, increase mixture components (stepwise to 8) and re-estimate.
- Step 6: Re-align.

[5] Unfortunately this system is not available for release. However, see Chapter 10 for the description and release of a tool which can produce very similar lattices for alignment.

- Step 7: Power normalise based on resulting alignment, repeat 1-6 before going to 8.
- Step 8: Trim silence above 50ms (both in alignment and wavs).
- Step 9: Re-initialise and estimate parameters based on silence trimmed alignment and wavs.
- Step 10: Repeat 3-6.

This is the method used in multisyn (Clark et al., 2004). However, similar procedures are generally used in TTS such as the EHMM procedure for festvox (Prahallad et al., 2006), which additionally allows for e.g., forward connected states in the HMM. While the proposed method in Prahallad et al. (2006) is arguably more suited for conversational speech, it is not employed as standard in neither the multisyn nor EHMM procedure and the focus in this chapter is on phonetic variation and not sub-phonetic variation.

5.3.2 Pronunciation Variant Alignment

The lattice approach proposed here, relies on two sources of information: hand-written variant options and pre-encoded pronunciation variants. The Combilex pronunciation dictionary (Richmond et al., 2009) contains pre-encoded pronunciation variants for a large portion of the dictionary, with the average number of variants per word being 1.82. However, the standard "full" pronunciation is what is normally retrieved (by Festival) from the dictionary for both the alignment procedure and at synthesis time, i.e., the variants are not utilised for training nor synthesis.

The lattice system first finds all variants present in the dictionary and populates the lattice, realised as a finite state transducer (FST), with them before expanding it using the additional manually produced rule-based variant options. These options are context-dependent rewrite rules written as regular expressions. They generally take the form of a phone with its left and right context and the resulting pronunciation variant, but can be any regular expression matching a valid string of symbols in the language of the Combilex dictionary. Thus the left and right context can be specific phones, features of the phone such as voicing, nasality and similar, but also word, syllable and phrase boundaries. The resulting variant option can be an alternative phone or a deletion of the phone. While the system can theoretically insert additional phones, this was not done due to the low number of insertions present in the gold standard compared to the Festival transcription (table 5.1).

The system can be run using either the Combilex encoded variants or a set of manually written rules, or a combination of both, as is explored in Section 5.4.3. The FST-based lattices are converted to the HTK standard lattice format (SLF) before use for alignment. Note that when using the lattices for alignment, they cannot be applied directly to the flat-start models in Section 5.3.1. Therefore step 1-3 proceeds as normal, however, at Step 4 the lattices are used instead of just short pause models and they are used throughout after that.

A set of manually derived rules was created (Table 5.2). These rules were found through the following procedure. The two annotators each created a list of the most salient changes observed while correcting the standard phonetisation. These lists were then supplemented by the differences found in Section 5.2.2 before being pruned by the annotators separately with their discussion partners (Julia Hirschberg and Korin Richmond). The two annotators then met with their respective lists and agreed upon a final set. When picking rules an eye was kept on how often the phenomena was present and whether the rule was similar to variations observed by other researchers. In fact, many of the proposed changes are similar to previously noted variants. For example in the Buckeye corpus (Pitt et al., 2007) end of word */t/* and */d/* are often deleted, but also word medial ones (Raymond et al., 2006). While we did not observe widespread word-medial */t/* and */d/* deletion, end of word deletion was observed and it was found that in the left context of a */s/* and */D/* they can be deleted. End of word stop deletion and glottalisation is also a common phenomena and has been often previously observed. One difference is that we do not take assimilation into account, although it is a common phenomena (Pitt and Johnson, 2003), however that was not often observed – one assimilation we do have, however, is voicing changes which some of were included. See Table 5.2 for the final set of rules implemented. Note that the rules may be speaker dependent as they are based on the differences in the 50 parallel read and spontaneous sentences. This is acceptable as it is of interest whether or not a relatively small manual effort can improve the alignment procedure in general.

5.3.3 Data

The read and spontaneous corpora from Chapter 3 were used. For each, the 50 sentences of the respective speech type from the gold standard corpus are also included.

Rule	**Description**
ANY < t WORD BOUNDARY > ANY → ?	Any end of word /t/ can be glottalised.
ANY < z > ANY → s	A /z/ in any context can be devoiced.
ANY < STOP WORD BOUNDARY > STOP → DELETION	Any stop at a word boundary followed by a stop can be deleted.
ANY < STOP WORD BOUNDARY > ANY → DELETION	Any word final stop can be deleted.
PHRASE BOUNDARY < VOWEL > ANY → REDUCTION	Any phrase-initial vowel can be reduced.
ANY < d WORD BOUNDARY > ANY → ?	Any word final /d/ can be glottal.
ANY < g WORD BOUNDARY > ANY → ?	Any word final /g/ can be glottal.
n < k WORD BOUNDARY > ANY → g	A /k/ after an /n/ before a word boundary can be voiced.
j < u > ANY → DELETION	A /u/ preceded by a /j/ can be deleted.
ANY < d > D → DELETION	A /d/ followed by a /D/ can be deleted.
ANY < t > s → DELETION	A /t/ followed by a /s/ can be deleted.
CONSONANT NO WORD BOUNDARY < I > CONSONANT NO WORD BOUNDARY → DELETION	A /I/ between two consonants in a word can be deleted.
f < @ > STOP → DELETION	A schwa after an /f/ and before a stop can be deleted.
ANY < VOICED CONSONANT > SENTENCE BOUNDARY → VOICELESS CONSONANT	Any voiced consonant at the end of a sentences can be devoiced.

Table 5.2: The hand-written pronunciation variant rules used.

Note this means 50 sentences are in the training data which also serve as test data. This is perfectly reasonable as we are not creating a phone recogniser, but rather attempting to create the best possible alignment of a *known* training set given no previous phone information. We are never trying to align unseen test data as the training data *is* the test data in our case.

5.4 System Comparisons

There are many factors to consider when performing forced alignment. In the following sections I will be comparing many different setups and variations.

In the following comparisons two systems lie at the core of the different variations:

- A standard system using the Festival multisyn tools and standard Festival pronunciation.
- A system using the Festival multisyn tools modified to run using lattices for pronujnciation variant alignment.

Each system, however, can be run in a variety of ways which affect the resulting alignment and in the following parts of this section I will go through a variety of different settings and their effects on the alignment accuracy.

5.4.1 Disregarding Minor Pauses

Firstly it is worth noting that what will here be described as the "standard system", differs slightly from the system used to produce the alignment which the annotator's corrected.

The system used to produce the output for correction used a triphone context model estimated based on a very large corpora of female RP English speakers for model initialisation. It used short pause modelling for the read speech and was provided pause information for the spontaneous. The large triphone model was initially created using a method similar to the standard procedure, and alignment using HTS' HSMMAlign method was done to obtain the triphone alignment.

The standard system used the method described in Section 5.3.1, which differs by using a monophone gaussian mixture model instead of the triphone model, was not initialised from models trained on data outwith the training data and was not provided accurate pause information for the spontaneous corpora.

The reason for this difference is due to the system used for correction being initially created for other purposes and was adapted for use here. The application of triphone modelling may seem like a better choice, however, in informal chats with other researchers it has been the generally stated opinion that monophone mixture models provided better alignment. The present author has however not found any citable evidence either way. It should be noted as well that during HMM modelling the alignment is actually slightly redone during the forward-backward algorithm and this will be in light of the full-context model application. Meaning that the main effect of the monophone vs. triphone initial model is at model initialisation.

Crucially, however, Festival is used for the phonetisation of the text, meaning the phone identity in both systems is identical except for differences in pause insertion. This difference occurs due to two factors. The fact that the corrected system initialised its models from a very large robust pre-trained model with good silence statistics, and due to the corrected system being given pausing information for the spontaneous speech from the orthographic transcription. The standard system uses flat-start initialisation and is not provided any prior[6] pausing information - a harder task.

Comparing the two systems' output in Table 5.3 we can see that the corrected system is slightly more accurate than the standard system. Most of the difference is due to additional short silence segments in the standard system, and, as the system used for manual editing was given the correct transcribed pauses, these additional pauses should all be incorrect. Therefore, an additional silence removal step was added to the procedure:

- Step 11: Remove silence below 40ms (in labels only) and re-estimate.
- Step 12: Perform final alignment using labels obtained from silence removal.

The threshold of 40ms was determined in part through the observation that in a 5-state HMM with a 5ms frame shift (used throughout in this thesis) anything less than 25ms will not be utilised at synthesis time (during training states can be skipped but not during synthesis) and in part through the observation that in the range 25-50ms (more than 50ms is already trimmed down to 50ms) 40 ms gave the highest accuracy. This step reduced the overall difference between the two systems by 3.4% (read) and 1.0% (spontaneous) absolute and was therefore employed in all the following systems, both

[6] During synthesis pauses are normally inserted at punctuation, however, during alignment only sentence inital and final pauses are required, all mid-sentential pauses are inferred using the short pause modelling.

	Deletions	**Insertions**	**Substitutions**	**Total**	**PER**
Read					
Standard	9	159	197	365	22.5%
Standard + pau fix	10	149	151	310	19.1%
Corrected	4	141	149	294	18.1%
Spontaneous					
Standard	16	218	181	415	26.2%
Standard + pau fix	17	202	180	399	25.2%
Corrected	14	187	175	376	23.7%

Table 5.3: The overall differences between the standard, standard with additional silence removal and the system used for manual correction (corrected). The annotator's agreed annotation is used as the reference in each case.

standard and lattice-based. Although they are not quite on par with the system used for correction, they are reasonably close considering the lack of prior silence knowledge.

Note that the systems shown here were run without phone reductions (see next section) as the corrected system did not employ any phone reductions.

5.4.2 Using Multisyn Phone Substitutions

The multisyn tools already allow for some variation in the alignment compared to the standard transcription produced by Festival, namely the option of including potential phone substitutions[7]. This has not been mentioned, or enabled, in the above discussion, but it has the potential to do similar things to many of the lattices, e.g., the manual rules contain the option of reducing all phrase initial vowels and many Combilex variants contain vowel reductions. However, it is possible that adding these additional potential reductions in combination with the systems presented here, could expand the lattices resulting in too many paths leading to reductions, which could bias the models toward too many reductions.

The "standard" set of phone reductions is fairly simple and allows any vowel to be reduced to a schwa. Table 5.4 shows the effect of enabling or disabling this. In all cases the phone reductions *increase* the overall error rate, particularly substitution

[7] This is implemented by having each phoneme as a "word" in an HTK dictionary which HTK then expands when recognising the phones. Whereas our lattices are produced from word-level variants creating lattices of these "words" for HTK to recognise

	Deletions	Insertions	Substitutions	Total	PER
Read					
Standard w Red	10	148	204	362	22.3%
Standard w/o Red	10	149	151	310	19.1%
Lattice w Red	22	106	144	272	16.7%
Lattice w/o Red	22	101	142	265	16.1%
Spontaneous					
Standard w Red	17	202	202	421	26.6%
Standard w/o Red	17	202	180	399	25.2%
Lattice w Red	39	128	149	316	19.9%
Lattice w/o Red	38	130	145	313	19.7%

Table 5.4: The effect of applying standard phone reductions. The annotator's agreed annotation is used as the reference in each case. Red = Reductions.

errors. This is, of course, due to additional schwa related errors which increase a lot (Table 5.5). It is interesting to note that even in the standard system this seems to create additional confusion, though possibly this is due to transcriber bias toward the initial systems output (see Section 5.4.4). Notably, however, when also using the lattices there is very little effect of applying or not applying the reductions. The lattices tend to lead to more schwa errors than the standard system with no reductions, and, just as when phone reductions are allowed, this comes primarily as additional substitutions. One potential explanation for this could be that when all, or most, vowels can be a schwa, the schwa HMM model becomes more of a general vowel model than a specific schwa model, and thus becomes generally plausible, i.e., it dominates the models.

Because not applying the phone reductions generally improved performance compared to the annotator's agreed transcription, these were not used in the subsequent systems.

5.4.3 The Effect of Lattices on Forced Alignment.

The lattice system can be run in three ways: hand-written rules only, the dictionary pronunciation variation only, or a combination of both. As each source of information may have a different effect, all three configurations were run. Using the agreed gold standard transcription as the comparison point, Table 5.6 lists the performance of each system. The proposed full system, i.e., lattices with both rules and Combilex variants,

	Deletions	Insertions	Ins Subs	Del Subs	Total
Read					
Standard w Red	1	15	82	18	116
Standard w/o Red	1	9	3	35	48
Lattice w Red	2	29	46	15	92
Lattice w/o Red	2	26	44	15	87
Spontaneous					
Standard w Red	0	18	46	14	78
Standard w/o Red	3	11	2	27	43
Lattice w Red	2	25	33	20	80
Lattice w/o Red	2	25	29	21	77

Table 5.5: Schwa related errors when applying standard phone reductions or not compared to the gold standard transcription. Red = Phone Reductions. Ins Subs = Substitutions where the schwa is incorrectly substituted into the sentence. Del Subs = Substitutions where the schwa is incorrectly replaced by another vowel.

improves the phone accuracy in both the read (2.8%) and spontaneous (5.4%) case compared to the standard method. However, just using the manual rules are better in both cases, improving the standard system by 3.9% and 6% absolute respectively.

Using lattices particularly reduces the number of phone insertions when comparing to the annotator's agreed transcription, meaning that using pronunciation variants are less likely to include phones not present in the audio, as compared to the traditional method. This is important as the cascading errors of Section 5.2.3 occur because of additional phones. For the systems using manual rules, substitution errors are also reduced whereas deletions increase slightly, though much less than the reduction in insertions. This suggests that the manual rules catch additional speaker specific variation not included in the dictionary.

While it seems clear that the manual rules reduce the number of errors, it is not clear that the Combilex pronunciation variants do. They do reduce the number of insertions but increase the number of substitutions. This is likely due to the double bias toward the standard transcription (described in Section 5.2.1), i.e. many of the variants in Combilex will contain substitutions, often dispreferred by the annotators, compared to the original transcription by the corrected system. Substitutions in particular are more debatable than deletions and thus likely to be more affected by this bias, and this bias

	Deletions	Insertions	Substitutions	Total	PER
Read					
Standard	10	149	151	310	19.1%
Lattice only dictionary	6	136	181	323	19.9%
Lattice only manual	20	106	120	246	15.2%
Lattice both	22	101	142	265	16.1%
Spontaneous					
Standard	17	202	180	399	25.2%
Lattice only dictionary	9	178	200	386	24.4%
Lattice only manual	37	133	134	304	19.2%
Lattice both	38	130	145	313	19.7%

Table 5.6: The effect of applying either the hand-written rules or the dictionary variants alone.

may have lead to the Combilex variant-based system seemingly not being useful.

5.4.4 Manual Alignment Bias

It is disappointing, and unexpected, that the Combilex-derived pronunciation variants do not provide a benefit in terms of reduced phone error rate. It is also unfortunate as the manually derived ruleset is not, obviously, automatic and that is what are interested in.

The explanation might be found in the doubly biased alignment toward the standard transcription (see Section 5.2.1). The manual rules are derived from differences between the gold standard and the standard method, so it is unsurprising that these rules are not affected by the bias, they essentially represent a specific set of changes from the standard to the gold standard. The Combilex variants are however not guaranteed to follow the same pattern, and describe, at least partly, a lot of other potential pronunciations, which were not addressed in the manual rules. These variants may, however, be perfectly valid and another annotator, not biased toward the standard transcription, may find that they agree with changes based on these.

To test the effect of the transcriber bias, and to see if the Combilex transcription could provide a benefit, a third annotator was asked to do an alignment correction, not from the standard automatic transcription, but rather from the output of the full system using both Combilex and manually derived variants. Using the output of the full system

	Deletions	**Insertions**	**Substitutions**	**Total**	**PER**
Read					
Corrected Automatic	15	169	179	363	22.6%
Annotator 1	72	84	132	288	17.9%
Annotator 2	60	81	133	274	17.0%
Consensus Trans	61	76	125	262	16.3%
Spontaneous					
Corrected Automatic	12	281	161	454	30.5%
Annotator 1	46	146	123	315	21.2%
Annotator 2	44	151	140	335	22.5%
Consensus Trans	46	142	131	319	21.4%

Table 5.7: The comparison of the third annotator to the others. Each comparison is to the third annotator's transcriptions. The corrected automatic system is the output of the system the original annotators corrected. The consensus trans(cription) is the agreed transcription between the initial annotators.

should bias the annotator toward this system, without directly biasing toward only the Combilex variants.

A third annotator was therefore recruited[8] and asked to do the annotation exactly like the first two annotators, except the initial transcription came from the full lattice system. To ensure any bias happened as an effect of annotation and not experimental bias, the third annotator was unaware of this discrepancy and only knew they were correcting the output of "a forced alignment system".

Table 5.7 shows a comparison of the third annotator to the other two, the gold standard alignment and the automatic standard alignment. As can be seen the third annotator is closest to the gold standard, but much further away than either of the two original transcribers (see Table 5.1). The standard automatic method is also further away than before, suggesting that starting from a different alignment has moved the annotator even further away from where we started. This amply demonstrates how much the initial starting point affects annotation!

Looking at Table 5.8 we see how the different systems presented in Table 5.6 perform when held up against the third annotator's transcription as the gold standard. Here we see a different pattern than before, now *both* Combilex pronunciations and manual

[8] Cassia Valentini-Botinhao of CSTR.

	Deletions	Insertions	Substitutions	Total	PER
Read					
Standard	15	169	179	363	22.6%
Lattice only dictionary	6	154	111	271	16.9%
Lattice only manual	10	111	102	223	13.9%
Lattice both	8	102	36	146	9.1%
Spontaneous					
Standard	12	281	161	454	30.5%
Lattice only dictionary	3	268	117	388	26.1%
Lattice only manual	9	201	100	310	20.8%
Lattice both	7	193	55	255	17.1%

Table 5.8: The effect of applying either the hand-written rules or the dictionary variants alone when comparing to the third annotator.

rules help, in particular with regards to substitutions. Manual rules still provide the largest benefit, however when combined the best results are obtained (9.1% read and 17.1% spontaneous). This suggests that the combination of both systems may actually be beneficial. As this test is biased toward the combined system, we cannot say for sure if the combination is better, in fact, if anything, what the results from the third annotator shows is that this is a problematic metric. In the next section a perceptual test in a real synthesis scenario is therefore reported.

5.5 Synthesis Evaluation

While an improved PER has been obtained, this might not translate into better synthesis quality. Furthermore, the discrepancy between annotator opinion and system PER casts doubt on the reliability of this as a measure of system performance. Therefore a synthesis-based evaluation was performed, i.e., a task-oriented evaluation of the resulting alignments (van Bael et al., 2007), which is also the ultimate goal of this work. Specifically, HTS 2.3beta (Zen et al., 2007a) was used to train eight HMM voices: four of each speech type, each using either standard alignment or one of the three lattice systems. The same corpora as for the alignment were used, i.e., the read and spontaneous corpora from Chapter 3. The 50 parallel read and spontaneous sentences which were included for the alignment corpora were excluded from voice training in

System	R-N	R-A	R-P	R-M	R-S	S-N	S-A	S-P	S-M	S-S
S-S	*	*	*	*	*	*	1	*	1	-
S-M	*	*	*	*	*	*	1	*	-	
S-P	*	*	*	*	*	*	*	-		
S-A	*	*	*	*	*	*	-			
S-N	1	*	*	*	*	-				
R-S	*	<0.005	*	<0.05	-					
R-M	*	1	<0.05	-						
R-P	*	<0.05	-							
R-A	*	-								
R-N	-									

Table 5.9: Adjusted p-values after Holm-Bonferroni correction for the Wilcoxon signed-rank test. * = p<0.001. Labels as in Figure 5.2.

all cases. They were used as test sentences.

5.5.1 Listening Test Design

A listening test based on MUSHRA (ITU, 2014) was run. This is similar to a MOS test, but allows side-by-side comparison of stimuli (same sentence but different systems) on a sliding scale from 1–100. Stimuli were unlabelled and ordered randomly. Natural versions of both read and spontaneous speech were included with the synthetic systems. Participants were asked to rate the stimuli according to how natural they sounded, with at least one stimulus at 100 and the others rated in relation to this. No designated reference sample was presented since each set included multiple natural productions (read and spontaneous). 32 paid participants were recruited and performed the experiment in a sound-insulated booth wearing Beyerdynamic DT770 PRO headphones. Each rated 15 sentences from either of two non-overlapping subsets randomly selected from the 50 test sentences, along with an initial practice sentence not included in the analysis. As each screen contained 10 parallel stimuli, participants took 2-2 1/2 minutes per slide, and the full experiment took approximately 30-40 minutes per participant. This amounts to 16 evaluations of each speech sample, for a total of 480 datapoints per system. All test materials are available at the thesis repository (Dall, 2017).

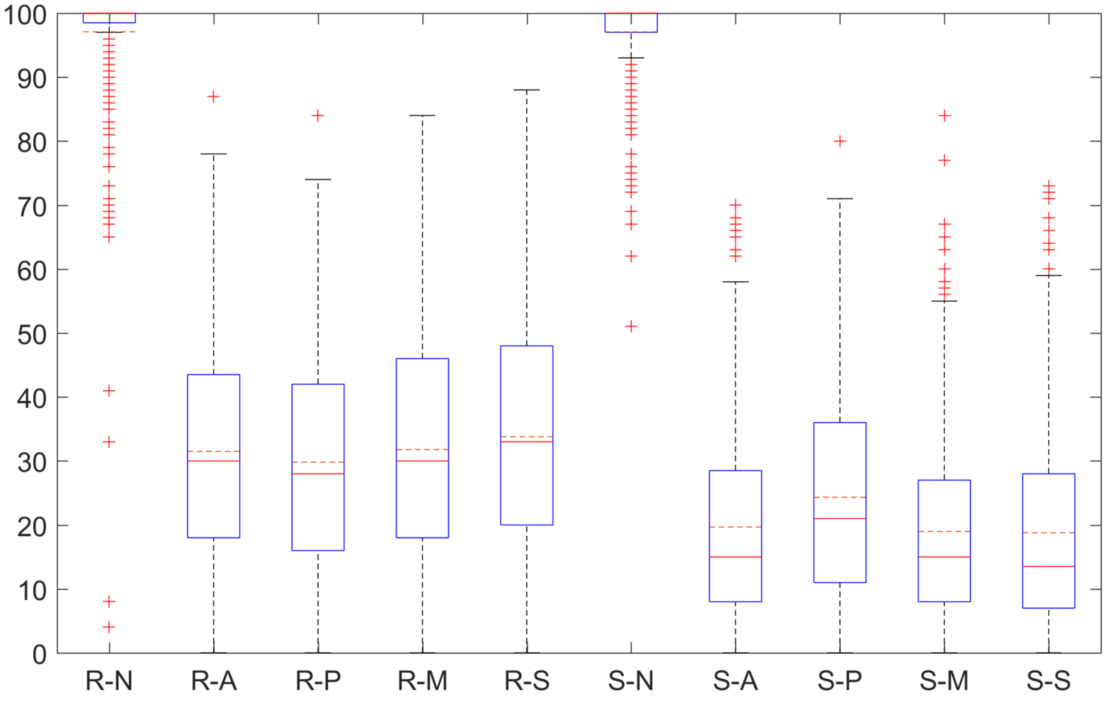


Figure 5.2: Boxplot of perceptual test results. R = Read, S = Spontaneous, N = Natural; M = Lattice w. Manual rules, P = Lattice w. Combilex variants, A = Lattice w. Both, S = Standard method. Solid lines are medians, stapled means and boxes 25 and 75% quantiles.

5.5.2 Results

The results of the test are graphed in Figure 5.2. Table 5.9 shows all system pairs compared using a Wilcoxon signed-rank test, after Holm-Bonferroni correction to avoid false positives. Natural speech is, unsurprisingly, rated significantly higher than synthetic speech. In contrast to the findings in Chapter 3, there was no significant difference between the natural speech types. However, this is probably due to the natural speech being clearly more natural than the synthetic, having the effect of minimising the perceived difference between the natural read and spontaneous sentences. All the read speech-based voices were rated significantly higher than the voices built on spontaneous speech. For the read speech, the standard alignment produces significantly higher rated speech than the other types, and the Combilex variants significantly lower, although the effect size is quite small. For spontaneous speech, on the other hand, the lattice system with only Combilex variants was rated significantly higher than all others, with no differences between the rest.

5.5.3 Discussion

The consistency assumption between training and synthesis only partly holds.

For read speech synthesis it seems to hold, as the standard method achieves higher ratings. However, informal listening to the output of the proposed pronunciation variant systems suggests that these systems produce hyper-articulated speech, which could reduce subjective naturalness. Arguably, however, we are getting what we ask for. At synthesis time we *ask for* the hyper-articulated version of the sentence, though we do not normally get it because of serendipitous reductions obtained as a result of the consistency assumption. Once we break that, a hyper-articulated version is produced. However, if we truly wish to control synthesis output, we should rather aim to have a better, more complete, acoustic model as provided by the proposed system. Methods for controllable, perhaps even gradeable, reduction of a sentence should then be developed, for instance utilising the reduced variants already encoded in dictionaries. While synthesis from the transcription found by the pronunciation variant alignments was not evaluated in a formal perceptual test, preliminary subjective evaluations are promising, indicating that pronunciation variant synthesis systems should be worth investigating by trying to mimic the most likely reductions produced by the speaker in the training data at synthesis time. See Chapter 6 for an exploration of both gradeable reduction and retaining of consistency between training and synthesis in a pronunciation variant setting.

Synthesis from spontaneous speech, in contrast to the read, appears to benefit from the use of pronunciation variants. It is encouraging that simply applying pre-encoded pronunciation variants, the Combilex only system, helps us learn a better model, particularly on difficult, spontaneous speech data, where a fully pronounced alignment is highly inappropriate. However, while including manual rules always improved accuracy, as measured by PER to the gold transcription, they did not increase perceived naturalness in synthesis. This may be due to them being overfitted to the test sentences, and thus not entirely suitable for the full training data. Equally, however, it could be due to these rules creating transcriptions which are so much further from the standard automatic ones, that they create such an inconsistency across training and synthesis, that although the models are better, they cannot be usefully used with the standard transcription. The fact that the read speech systems using the manual rules were not considered less natural than the Combilex system does not seem to support this conclusion, with both of them slightly better than the Combilex system. However, while the manual rules could theoretically move the read speech just as far apart as for the spontaneous, the spontaneous speech, in general, is further removed from the standard transcription, it is unlikely that this would happen in practice. In these experiments, this is also the case with the manual rules providing a smaller PER improvement (3.9% to gold, 8.7% to third annotator) than the spontaneous (6.0% to gold, 9.7% to third annotator). It is therefore quite possible that speaker-mimicking transcriptions can provide a way to retain consistency for the spontaneous speech even when applying speaker specific rules, something which is explored in Chapter 6.

Synthesis models from read speech were rated more natural than models from spontaneous speech. This is not surprising since spontaneous data is much more varied and difficult to model, as exemplified by the much less accurate alignments. However, using pronunciation variant forced alignment pushed spontaneous speech towards closing the gap, and likely forms part of the alignment method used by Andersson (2013). It is also possible that the difference found between read and spontaneous speech is due to difference in phone accuracy and not specifically the type of speech, i.e. that read speech with a lower accuracy can benefit from pronunciation variant forced alignment. This will not be further investigated in this thesis.

Notably improved accuracy in the phone modelling has improved spontaneous speech synthesis – confirming that one of the main issues with spontaneous speech is bad modelling as suggested in Chapter 3. With this improved acoustic model we can then investigate an issue more related to the prosodic issues with TTS – namely pronunciation variation, which we will investigate in the following chapter.

5.6 Chapter Conclusions

I have reported on an investigation into using pronunciation variants during forced alignment for read and spontaneous speech corpora in TTS. This has also tested the consistency assumption between training and synthesis. A pronunciation variant-based forced-alignment method was implemented using pre-encoded dictionary variants and a set of manually derived variant rules, and its application to speech synthesis evaluated. It was found that standard synthesis with read speech did not benefit from these variants, though the underlying acoustic model arguably was more correct. For spontaneous speech, including pronunciation variation yielded an improvement only for the tightly constrained dictionary encoded variants. This suggests that the consistency assumption between training and synthesis only holds when either minor errors in training transcription are made or when the consistency is only somewhat “broken”. From this, it is suggested that further improvements could be attained by incorporating automatic pronunciation reduction at synthesis time, something which will allow for the improvement of alignment transcriptions while also retaining consistency, and this will be explored in the next chapter.

Chapter 6

Pronunciation Choice for Pronunciation Variant Synthesis

A Note on Collaborative Work

The work presented in this chapter was not performed in direct collaboration with anybody. Part of the work was done while visiting Nagoya Institute of Technology and Keiichi Tokuda and Kei Hashimoto provided helpful initial discussion.

6.1 Introduction

In the previous chapter, it was tested whether breaking the consistency assumption across training and synthesis with regards to the use of pronunciation variant forced alignment is beneficial. While the underlying acoustic model became more accurate, it did not result in a tangible improvement in output synthesis quality – particularly for read speech-based voices. It was suggested that this was due to the tendency of the resulting voices to produce hyper-articulated pronunciations, which is arguably more correct as the phonetisation consisted of full pronunciations.

This chapter will focus on using language models to predict reduced pronunciations at synthesis time and also to retain consistency across training and synthesis (which was broken in the previous chapter) when choosing pronunciations.

Producing more reduced pronunciations should lead to speech more similar to that occurring spontaneously due to the tendency for spontaneous speech to be more reduced than read speech. This observation can be derived from the previous chapter's investigation into pronunciation variant forced alignment. One of the observations made about the difference between the two types of speech when manually correcting the standard alignment was exactly that: the spontaneous was more reduced. This was the case both in terms of phone deletions and reductions, particularly vowel reductions. This is not a new observation and has been noted by other researchers (e.g., Werner et al. (2004); Brognaux et al. (2014b)).

Furthermore, in a series of experiments intended to allow for durational control, with a specific view to produce faster speaking rates similar to those found in spontaneous speech, Werner et. al. (Eichner et al., 2003; Werner et al., 2004; Werner and Hoffmann, 2006, 2007) found that including pronunciation variants could be used to indirectly control speaking rate by choosing shorter pronunciations (in terms of number of phones). This not only allowed for their desired speaking rate control, but also to an improved perception of the resulting speech as reported in Werner and Hoffmann (2007). Shorter pronunciations are generally reduced pronunciations, and as such Werner et. al.'s findings are indicative that applying pronunciation reduction should not only lead to speaking rates which match spontaneous speech better, but also to improved perceived quality of the resulting synthesis.

It would seem that a key to achieve better, and more conversational, speech synthesis could be to produce such reduced pronunciations at synthesis time. Doing this may also restore the, in the previous chapter broken, consistency between training and synthesis by producing pronunciations similar to those found during pronunciation variant training. In fact, it is entirely possible that a voice based on pronunciation variant alignment is better suited to take advantage of such pronunciation variant synthesis due to its more accurate modelling of individual phones.

6.2 N-gram-based Pronunciation Reduction

Jurafsky et al. (2000) found that more probable words are more likely to be reduced. This observation was based on the use of various measures of word likelihood to derive the probability of each word in a sentence, in context. The most notable used method is that of conditional probability given previous, following or surrounding word context – essentially equivalent to a standard n-gram model (previous words), a backward n-gram (following words) or a centered n-gram model (surrounding words). Their definition of reduction included vowel reduction, end of word /t/ or /d/ deletion and durational shortening – both vowel reduction and end of word /t/ or /d/ deletion were amongst the most common differences between standard and pronunciation variant alignment found in the previous chapter – suggesting that n-gram language models are suitable for predicting word pronunciation reduction.

The basic idea is to take an n-gram language model (LM) and score each word in an utterance and then produce reduced pronunciations, if possible, of the words with the highest probability. In order to do this, a method for determining *what* a reduced pronunciation of a word *is* is needed, and also some measure of how many of the words in a sentence should be reduced. While one could train a letter-to-sound (LTS) model which could output alternative pronunciations and pick shorter pronunciations (in the style of Werner and Hoffmann (2007)), this method does not guarantee a pronunciation is reduced, nor correct. An immediately simpler approach, which is also likely to be more accurate, is to utilise the pre-encoded reduced pronunciation variants in Combilex (Richmond et al., 2009). In Combilex, each entry is marked as either "full" or "reduced" and, typically, if a pronunciation variant exist it will be a reduced variant of the full pronunciation. This has the added benefit of being more compatible with voices trained using the previously presented pronunciation variant forced alignment and should thus be able to retain consistency better.

In this approach, the important thing to determine is then whether producing reduced pronunciations results in better synthesis, and how much reduction produces the best/most natural pronunciations.

The standard method for choosing between pronunciations, utilised by most available open-source front-ends (e.g., Flite/Festival, Idlak, MaryTTS – see Chapter 8 for a description of each), is to simply pick the first encountered full pronunciation (sometimes just the first encountered pronunciation), and, in the case of Festival, also taking part-of-speech tags into account, and use that. This means that the "standard" pronunciation is generally the full pronunciation, and as mentioned several times, this could lead to the hyper-articulation observed with better alignment.

In the proposed method, a "reduction parameter" is used to choose the amount of reduction. Specifically, a value from 0 to 1 will be used where 0 is fully reduced and 1 is fully pronounced (and equal to the standard pronunciation selection). By setting the value to e.g. 0.5, results in the 50% most likely words in a given sentence, according to the LM, being attempted reduced. To do the reduction, each word is looked up in the dictionary and if a reduced version is available it will be selected. If multiple reduced variants exist, the first encountered will be chosen. This is a simple method, but a good choice for an initial investigation. One potential issue with this is that it may choose bad pronunciations, and of particular relevance to this investigation, it may select pronunciations which do not match those found during alignment, even though that pronunciation may be available.

In order to try and match the pronunciations found during alignment, a phone level approach is also investigated. In this approach, a phone n-gram is trained on the aligned corpus from pronunciation variant forced alignment, and is then used to decode a lattice containing all potential pronunciations also given to the forced alignment. The highest scoring path is then chosen as the pronunciation. In this way the best paths should be chosen to conform to the pronunciations produced by the alignment - retaining consistency across training and synthesis.

6.2.1 Phonetic Accuracy of Proposed Methods

To get an indication of the performance of each of the different models, in terms of proximity to the found pronunciations of the variant alignment, the same 50 sentences used for comparisons in the previous chapter were phonetised using each method. These can then be compared, using the same Levehnstein distance metric as in the previous chapter, to the found phonetisation from the alignment. Note, that this only tells us if we produce phonetisations similar to those found during alignment – not whether the found phonetisations will sound better. However, following Brognaux et al. (2014b) if hand-corrected pronunciations give better results, it is highly likely the found pronunciation variant alignment also does – i.e., through consistency and possibly simply better pronunciation choice.

The word level n-gram model was a model created for ASR and consisted of 330 million word tokens drawn from a variety of sources but most notably included a large number of TED talks – giving the LM some spoken sources in its training data (although of course TED talks are far from spontaneous). The model trained was a forward 3-gram model and it was used to predict word reductions with the reduction parameter set to 1 (standard, full pronunciation), 0.75 (25% reduction), 0.5 (half reduction), 0.25 (75% reduction) and 0 (fully reduced). The words the were reduced were those with the highest n-gram probability.

The phone level n-gram model was trained on each of the Read and Spontaneous data sets respectively and was a forward 4-gram model. As it was quite possible that phones in boundary contexts would behave differently than others, word boundary and syllable boundary information was included as part of the n-gram token string. A lattice of all possible pronunciations in Combilex for each word in a given sentence

	Deletions	**Insertions**	**Substitutions**	**Total**	**PER**
Read					
Standard	4	13	114	131	7.5%
25% Reduced	6	13	85	104	6.0%
50% Reduced	7	11	77	95	5.4%
75% Reduced	9	10	71	90	5.2%
Fully Reduced	9	10	70	89	5.1%
Phone LM	15	6	73	94	5.4%
Spontaneous					
Standard	0	17	93	110	6.3%
25% Reduced	2	17	79	98	5.6%
50% Reduced	4	16	87	107	6.2%
75% Reduced	5	14	88	107	6.2%
Fully Reduced	5	13	89	108	6.2%
Phone LM	10	9	89	108	6.2%

Table 6.1: Differences between the pronunciation found by variant alignment based on Combilex to each of the different pronunciation methods. PER = Phone Error Rate.

was then created and decoded using the phone n-gram model.

Table 6.1 shows the accuracy of each method of pronunciation choice compared to the alignment found using pronunciation variant forced alignment for each speech type. As can be seen, in each case, the standard method of pronunciation selection is the furthest from the found alignment. For the read speech each increase in the level of reduction prediction further decreases the discrepancy, though 75% and full reduction are basically equivalent. The phone n-gram method reduces the discrepancy as well, but not as well as the word-based reduction, albeit the differences are small. For the spontaneous speech, only the 25% reduction stands out slightly, with the rest basically equivalent.

However, this does not mean each method produces the same phonetisation. Table 6.2 shows each pronunciation selection method compared to the standard approach where we can observe that the differences get progressively larger, as opposed to the stagnation when compared to the alignment.[1] The progression is also natural for the

[1] The astute reader will have noticed that while the word level pronunciations would be the same for spontaneous and read speech, the phone level should not. However, after double checking the result, the present author can confirm that this was indeed the case - despite the phone level models being different.

	Deletions	Insertions	Substitutions	Total	PER
25% Reduced	2	0	68	70	3.9%
50% Reduced	5	0	109	114	6.4%
75% Reduced	8	0	138	146	8.2%
Fully Reduced	8	0	144	152	8.5%
Phone LM	20	2	149	171	9.6%

Table 6.2: Differences between the standard pronunciation and the different pronunciation methods. The shown numbers are the same for both read and spontaneous speech. PER = Phone Error Rate.

word-based reduction as each will produce progressively further reduced words. This suggests that while the different methods might find different pronunciations, they all find some part of the pronunciation also found by the alignment.

In general the major difference between the standard phonetisation and the alternatives is the amount of vowels reduced to /@/, which progressively increases. For the phone LM-based phonetisation a number of deletions additionally occur which is primarily end of word /d/ deletions. Both these phenomena are part of the major differences between the standard and manual phonetisations found in the previous chapter.

These results suggest that both word and phone level pronunciation prediction are useful for more accurately predicting the pronunciations found to have been actually produced by the voice talent. This does not, however, guarantee that these actually sound better, although as previously noted the results of Brognaux et al. (2014b) suggests that this should be the case. Therefore a listening test was run.

6.3 Perceptual Effect of Pronunciation Reduction

A MUSHRA-style test similar to that in the previous chapter was carried out, the major difference being the decision not to include any natural reference sentences such that the test only contained synthetic speech. The decision to do this was made based on the observation that the inclusion of the natural references in the previous chapter seemed to squash the perceived differences between the synthetic voices, and this would be desirable to avoid. Furthermore, the test was run separately for each speech type, such that participants only compared systems of one speech type to each other

The reason for this is probably due to /@/ dominance.

and not between speech types. This decision was made after observing that spontaneous speech-based voices do not sound as good as read speech-based voices even after pronunciation variant forced alignment, and the informal observation that this gap is not bridged by pronunciation reduction.

For each speech type – read and spontaneous – the voice based on the Combilex derived pronunciation variant forced alignment from Chapter 5 was used. The choice of using these voices was twofold. Firstly, for the spontaneous speech it was the best voice. Secondly, although the Combilex-based voice was the worst rated of the read speech-based ones, we are here utilising only Combilex-based pronunciations and so to ensure a higher level of consistency this voice should produce better results - see Section 6.4 for more on this.

A subset of 21 randomly chosen sentences of the 50 test sentences from the previous chapter was used as the test material. Each test sentence was synthesised in 5 different ways – standard pronunciation, half reduction, full reduction, phone LM pronunciation and alignment found pronunciation. The pronunciation found by alignment based on Combilex encoded variants was included for two reasons. Firstly, to see if such a pronunciation sounds better in general. Secondly, assuming it does, to provide a measure of the potential gain from better pronunciation choice – though of course this pronunciation is still limited to which pronunciations are available in the lexicon (more on this in Section 6.5). Although half reduction did not provide the best accuracy compared to the alignment, it was included in the test as there were fears that constantly reducing would be too much, and therefore a more measured approach might turn out to be better. 75% reduction was also an option, but was deemed too similar to the full reduction.

30 paid native speakers of English were recruited to take part in the study. Each participant first rated all spontaneous voice based samples and then the read voice samples. Each participant was provided with one sample sentence to accommodate themselves to the experimental setup. This was followed by the 20 remaining sentences. Participants were asked to rate how natural each system sounded from 0-100 without further qualification on what natural might mean. If participants asked what natural meant, they were told to apply their own judgement. As mentioned previously no natural reference was included.

All samples and experimental materials available at the thesis repository (Dall, 2017).

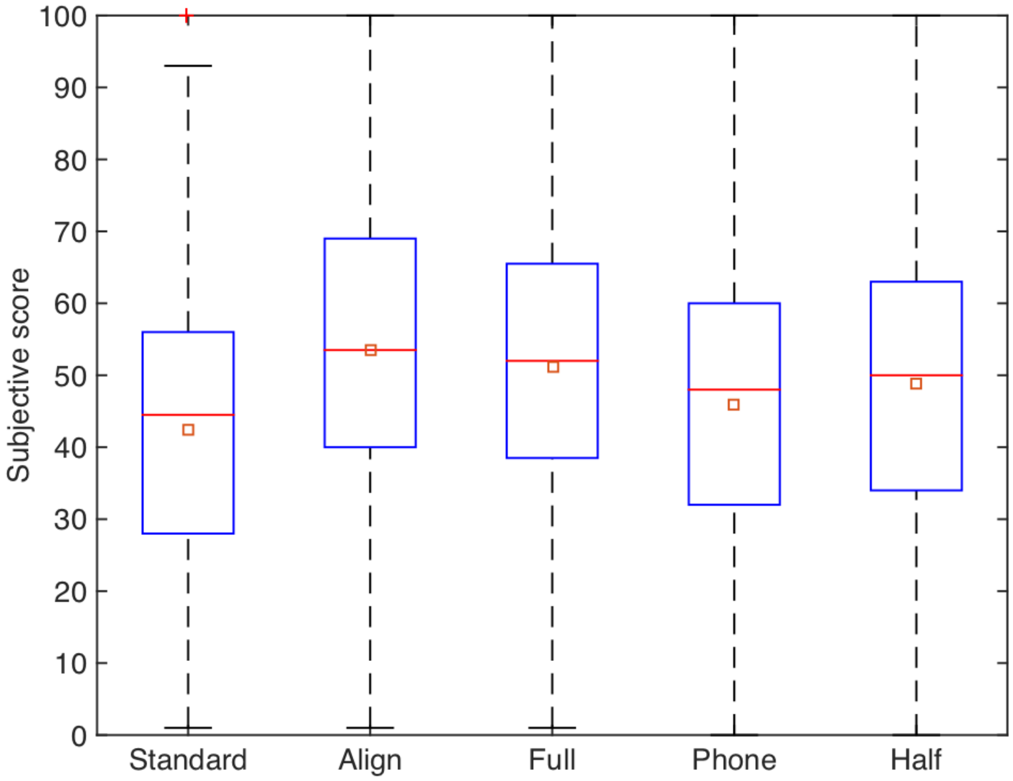


Figure 6.1: Boxplot of perceptual test results for the read speech-based voices. Standard = Standard pronunciation. Align = Alignment-based. Full = Full reduction. Phone = Phone LM-based pronunciations. Half = Half reduction. Solid lines are medians, squares means and boxes 25 and 75% quartiles.

6.3.1 Results and Discussion

One participant did not finish all slides and consequently was excluded from the analysis. Figure 6.1 shows the results for the read speech-based voice. Using Wilcoxon Signed Rank test for statistical significance, and after applying Bonferroni correction, all pronunciation methods are statistically significantly different at least at the $p<0.05$ level. That is, the standard pronunciation is considered the least natural, than the phone LM, half reduction, full reduction and finally the found phonetisation by the Combilex-based alignment. For the spontaneous speech, Figure 6.2 shows the results. Using the same statistical significance test, it was found that there was no significant preference for the phone LM pronunciation over the standard pronunciation, however both where statistically significantly dispreferred compared to the others. There was also no difference between half reduction and full reduction, and no difference between

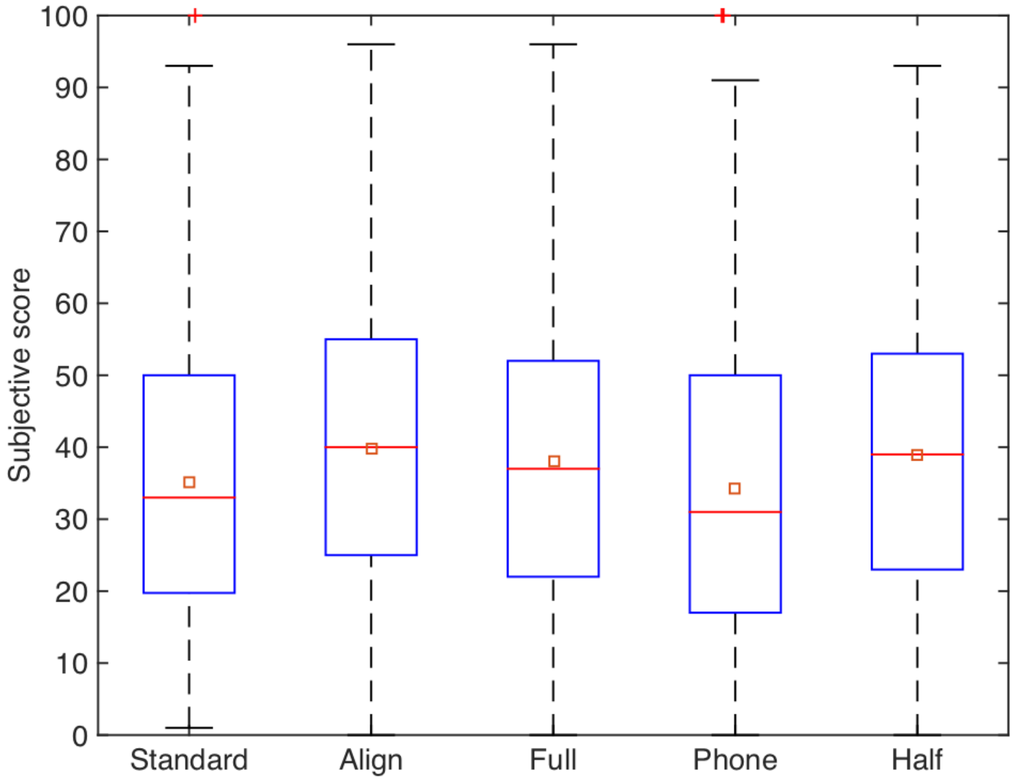


Figure 6.2: Boxplot of perceptual test results for the spontaneous speech-based voices. Standard = Standard pronunciation. Align = Alignment-based. Full = Full reduction. Phone = Phone LM-based pronunciations. Half = Half reduction. Solid lines are medians, squares means and boxes 25 and 75% quartiles.

half reduction and alignment found pronunciation. However, the pronunciation found by the alignment was significantly preferred over the full reduction system. In other words, there seem to be two groups, the standard and phone LM-based in one group, and the alignment found, half and full reduction methods in another. Within the second group perceptual results are closer,, particularly for spontaneous speech, however the tendency is that the alignment found is preferred, followed by half reduction and finally the full reduction system.

The overall tendency for both types of speech-based voice is thus that the alignment-based pronunciation is considered more natural than the others. This is in line with the findings of Werner and Hoffmann (2007), Brognaux et al. (2014b) and also the consistency assumption. The alignment found pronunciation is consistent as it follows the same pattern as that of the pronunciation variant alignment – which the standard pronunciation does not.

Word level n-gram reduction prediction in either the full or half setting also provides a clear benefit over standard pronunciations for both types of speech. Whereas the phone LM prediction did not seem to be beneficial for the spontaneous speech, it has some merit for the read speech. With regard to consistency, the phone LM pronunciation should theoretically be the better of the non-alignment-based methods as it infers its probabilities directly from the found alignment. However, the reason that it does not match the alignment-based pronunciation can probably be found in the local nature of the n-gram's decisions and the pervasiveness of /@/. The schwa is the primary phonetic difference found between the alignment and the standard pronunciation, with many more /@/ found in the alignment. Combined with the fact that /@/ is already the most common phone has likely lead the /@/ model to dominate, by being very probable in any context, the n-gram LM. This is evident with the number of /@/ predicted by the phone LM similar to that predicted by full reduction.

The results obtained using full and half reduction suggest that, even in read speech, words are seldom fully pronounced and that full pronunciation is seen as unnatural to a certain degree. This benefit is predicted by the results in Chapter 3 where it was shown that naturally produced spontaneous speech is found to be more natural than read speech, and these reduced pronunciations are likely more towards the conversational spectrum than the read. In particular, the full reduction system in the read speech case, has likely benefitted from this, which can explain why the half reduction tends to be considered more natural for the spontaneous speech; the read speech with full reductions becomes more like spontaneous speech, whereas spontaneous speech with full reduction is already very spontaneous and perhaps goes a bit too far. It has to be noted, from informal listening, that the spontaneous speech-based voice is still very inconsistent and it is possible that artefacts have influenced the results to a larger degree.

It is also clear that the spontaneous speech-based voice is still inferior to the read speech-based one. While the two MUSHRA-style tests presented here cannot be directly compared, participants did the read speech-based test after the spontaneous, and the fact that subjects generally rated read higher than spontaneous could be a result of an indirect comparison. Informally listening to any of the test samples in the thesis repository leaves little doubt that the read speech-based voice is the better system. The question as to why this may be, particularly when using the correct gold standard pronunciation, is important. If one listens to sample 1 from the MUSHRA tests of both the Read and Spontaneous voices it is clear that the main problem is not so much prosodic as it is with the general modelling (although it certainly has a prosodic element). Listening to the sample one can notice that for several words a crackling buzz appears (e.g. "like" and "applied") in the spontaneous samples which are not in the read. This crackling is problematic and could be due to further issues with general modelling of the speech, both in terms of pronunciation and alignment, but another possibility is that of the speaker using a creaky voice at times which is badly modelled by current vocoders (this is evident in the natural speech samples of the parallel read and spontanoues where creakiness appears in the spontaneous and not the read. This is an issue which cannot be overcome in this thesis, improving vocoding is beyond the current scope, and in order to focus on filled pause modelling we will turn our attention to read speech for the rest of this chapter. See Chapter 7 for a discussion of how spontaneous speech may still be utilised to synthesise more natural filled pauses.

6.4 Testing the Consistency Assumption (Again)

The above results indicate that the Combilex pre-encoded reduced variants are in fact better for general synthesis than the full pronunciation variants. This is particularly true for the read speech-based voice which sees the largest improvement. However, the above test was done using the voice based on pronunciation variant forced alignment, and that was rated worse than a standard voice in the previous test (in Chapter 5) using read speech. Considering that simply applying reduction when possible can also be done with a non-variant alignment-based voice, it is important to see whether such a voice benefits in a similar manner.

It is the hypothesis here that although the non-variant voice should also be able to benefit from better pronunciation choice, the variant voice should gain a larger benefit as it will better realise the correct phone string as compared to the non-variant. This is supported by the observation in the previous chapter that the variant voice was rated slightly less natural, however, it tended to hyper-articulate. Just as the non-variant voice would benefit from "accidental" reductions, it should also see "accidental" hyper-articulation the other way. Furthermore, in the previous section the spontaneous speech-based voice also benefitted from pronunciation reduction, and this voice was based on the highest rated alignment system – suggesting this improvement is to some degree independent of the underlying alignment (or tied to pronunciation variant forced alignment).

To determine this, another MUSHRA-style test was done using just read speech-based voices. A voice based on the standard non-variant alignment and the same voice based on the Combilex pronunciation variant forced alignment was used. 5 different pronunciations were tested. Two from the standard voice, and three from the variant voice. From the standard voice the standard pronunciation which is consistent with the non-variant alignment (henceforth S_S) and the fully reduced pronunciation which is inconsistent with the non-variant alignment (henceforth S_F) was used. From the variant voice the standard pronunciation which is inconsistent with the variant alignment (henceforth V_S), the alignment found pronunciation which is consistent (henceforth V_A) and the fully reduced pronunciation (henceforth V_F) which is technically inconsistent, but closer to the variant alignment than standard pronunciation and more spontaneous in style, was used.

The choice of using the fully reduced version over the half reduced (or potentially 75% reduced) was partly motivated by the above reasoning that it should be more spontaneous, partly due to it being the superior method in the previous test and finally because it is the simpler method – one can forego the use of an n-gram model entirely and simply rely on improved, reduced, pronunciations in the dictionary (and by extension in a trained G2P model).

The test setup was exactly as above, except for the different systems, and 38 paid native English speaking participants were recruited to take part. All experimental materials and samples can be found at the thesis repository at (Dall, 2017).

6.4.1 Results and Discussion

Two participants were excluded from analysis on the grounds that they were clearly not native speakers of English.[2] One participants' results were incomplete and was thus also excluded from analysis. Figure 6.3 shows the results. Using Wilcoxon Signed Rank test for statistical significance, and after applying Bonferroni correction, all voices except V_F and S_F are statistically significantly different at least at the $p<0.05$ level. That is, using standard or pronunciation variant alignment with full pronunciation reduction is not considered more or less natural than each other. However, both methods are considered more natural then standard alignment with standard pronunciation, which in turn is more natural than variant alignment with standard pronunciation. Using variant alignment with alignment found pronunciation is the most

[2] Participants self-report nativity.

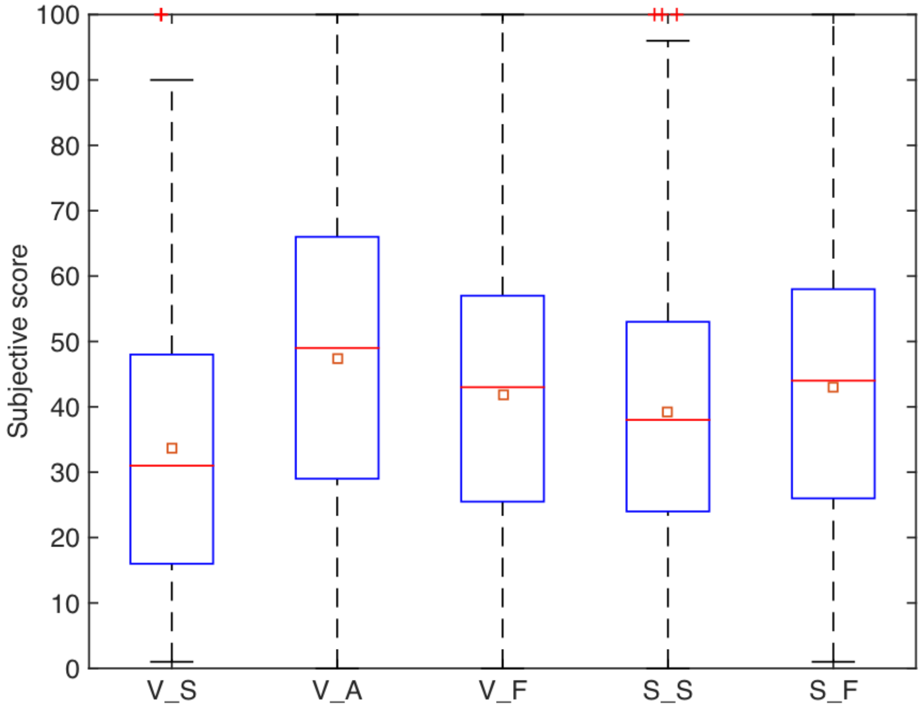


Figure 6.3: Boxplot of perceptual test results. S_S = Standard alignment, standard pronunciation. S_F = Standard alignment, full reduction. V_S = Variant alignment, standard pronunciation. V_A = Variant alignment, alignment-based pronunciation. V_F = Variant alignment, full reduction. Solid lines are medians, squares means and boxes 25 and 75% quartiles.

natural of all.

The results show that even when using a standard non-variant alignment-based voice some gain can be had from reduced pronunciation choice. In fact, using variant alignment does not provide any additional benefit in this case. However, the range of different perceptions of the variant alignment-based voice is notable. The standard non-variant alignment-based voice is much less receptive to changes in pronunciation, whereas the variant alignment-based voice output synthesis changes much more in reaction to different pronunciation choices. This is notable partly because the alignment found pronunciations highlight how better pronunciations can really improve a voice, and partly because it shows that a standard alignment voice is not suitable for showing the effects of, nor utilising, pronunciation choice effectively. That the alignment found pronunciations are rated much higher, highlights that there is significant potential gain in efficient pronunciation choice and that this is an area which should be further investigated – particularly in light of the hyper-articulation of the standard pronunciation.

The consistency assumption seems to hold with regard to pronunciation choice as well, the standard voice, while less reactive to pronunciation choice, produces very consistent pronunciations – almost regardless of pronunciation choice. When utilising a variant alignment the best results are obtained when using a pronunciation choice derived from the same alignment and using a reduced pronunciation closer to that improves naturalness as well – both lending support to the consistency assumption. This would suggest that further improvements could be obtained by training a voice in which the full reduction pronunciation is used for alignment *and* synthesis – thereby using better pronunciations and retaining perfect consistency. This will however not be investigated in this thesis. It can be informally noted that when the full reduction pronunciation method with variant alignment does not perform well is when it picks odd pronunciations – e.g. for the word “do” which sounds out of place in several instances (e.g., sample 08 from the experiment in the thesis repository). The non-variant voice using this pronunciation sounds less problematic as it manages to compensate slightly through it being less receptive to pronunciation change. This issue also highlights that fully reducing is a simple, effective, method – but that it is far from perfect. Further research into pronunciation choice could thus be a fruitful avenue of research – although it is considered out of scope in this thesis.

6.5 Chapter Conclusions

In this chapter two methods of pronunciation prediction have been proposed and compared to the standard method of pronunciation choice and also the effect of different alignments on the effectiveness on pronunciation choice has been investigated. It was shown that a utilising a word-level n-gram language model can be used to predict pronunciation reduction and that this can improve the perceived naturalness of a voice based on pronunciation variant forced alignment. A phone-level n-gram method was also proposed, and although it did improve upon standard pronunciation choice it was not as effective as word-level reduction prediction. It was furthermore shown that while this had a positive effect in both read and spontaneous speech-based voices, the spontaneous speech-based voices did not improve to a point where they could be considered on par with a read speech-based voice. It was therefore decided not to further investigate voices based directly on spontaneous speech – and in the next chapter we will turn our focus into utilising data mixing techniques for filled pause synthesis.

Furthermore, the initial experiment relied on a pronunciation variant forced alignment, however the highest rated of the proposed methods is simply reducing when possible – this can also be done using a voice based on standard alignment. In a second experiment it was tested whether a standard alignment-based voice could also benefit from fully reduced pronunciations. It was shown that this indeed was the case, however it was also shown that such a voice was less reactive to pronunciation choice and that it is unlikely that a standard alignment-based voice would be able to benefit from better pronunciation choice to the same degree as one based on variant alignment. It was suggested that research into pronunciation choice could be a fruitful endeavour due to the high rating of the alignment-based pronunciation choice and that this is best done using variant alignment-based voices. This is, however, considered out of scope of this thesis.

Chapter 7

Synthesising Filled Pauses

A Note About Collaborative Work

The work in this chapter was performed in collaboration with Mirjam Wester (CSTR) and Marcus Tomalin (Cambridge) who were both invaluable discussion partners throughout. Marcus specifically also helped find useable sentences for the perceptual tests in the Desert Island Discs material.

7.1 Introduction

In the previous chapters, the focus has been on improving the base quality of a synthetic voice based on read or spontaneous speech. In this chapter, the focus will turn to synthesising filled pauses (FPs) – specifically 'UH's and 'UM's.

In Chapter 4, it was shown that some of the psycholinguistic effects of FPs in naturally produced speech did not occur when using synthetic speech. The primary suspected reason for this is poor FP realisation by the TTS system – presumably due to lack of training data. The most straightforward solution would thus seem to be training a spontaneous speech-based voice. However, as seen in chapter 3, and still after the improvements in Chapters 5 and 6, the naturalness of such voices lags behind that of read speech-based voices, to an extent where it is unclear whether improved FP synthesis will be overshadowed by decreased naturalness. Andersson (2013) found that the inclusion of FPs increases the naturalness of synthetic speech based on a spontaneous voice, when compared to one based on read speech – this claim was investigated in Chapter 3 and shown not to hold for the data collected here. However, based on the results of that test and the experiences of previous researchers, a method, inspired by Adell et al. (2012), based on the direct modelling of FPs using a specific phonetic representation, is presented. In this method, the phones /V/, /@/ and two non-standard representations based on these are compared. This allows for a direct and separate modelling of the vowels in FPs and their surrounding phones in context. Finally, a variety of data mixing techniques are investigated with the goal of retaining read speech voice quality while still synthesising convincing FPs. The techniques are compared to those of Andersson (2013) and include the direct mixing of spontaneous and read speech corpora, corpus marking (the technique of Andersson (2013)), speaker adaptation techniques and also a mark switching technique developed in this chapter.

In general, the results of the various methods will be determined by preference tests, but the general acoustic notes from Chapter 4 will also be taken into account, and it is worth summarising here that the main effects of FPs are those of durational lengthening compared to standard phone durations and a fall in F0 compared to surrounding phones.

7.2 Phonetic Representation of Filled Pauses

As the improvements found by Adell et al. (2010a) came from a specific FP model, something similar is proposed, although in an HMM-based framework. The proposed model of Adell et al. (2010a) relies on an analysis of the acoustic features of FPs, such as increased duration and lowered F0, and was required to be explicit in order to guide the unit selection directly. However, in an HMM-based framework we are already building models of each context-dependent phone, and therefore we do not need an explicit FP model. Instead, we need the linguistic context features of the phones in an FP to distinguish it from other phones of the same type, which would allow the decision tree context-clustering to group the FP phones together. If a high-quality part-of-speech tager is used, FPs should be tagged as such, and this could serve as the distinguishing feature. However, FPs are usually not well modelled by sentence structure POS-taggers, and, in standard front-ends such as Festival and Flite, the tag set is reduced to one that does not contain the FP tag. In Festival 2.4 (Black et al., 2014), when using the Combilex dictionary (Richmond et al., 2009), an ‘UH’ is phonetised as /V/ – an unrounded, open-mid, back vowel – and an ‘UM’ as /V m/. However, for ‘UM’ there are two additional alternatives in the dictionary: /@ m/ – schwa followed by a bilabial nasal - and /m!/ – a short bilabial nasal. While these are never used in standard transcriptions, they provide a convenient alternative representation. We here ignore the reduced form of /m!/, partly because it could arguably be considered to be the backchannel 'mhm' and not an FP, and partly because we are initially interested in fully pronounced FPs and not heavily reduced versions. Using /@/ or /V/ for both 'UH' and 'UM' does not, however, uniquely identify FPs as the /@/ is the most commonly occuring phone and /V/ is also present in other words. /@/ may however be a better representation of the vowel of an FP than /V/, and so should be considered. In order to provide a distinguishing feature, it is here proposed that a separate phone identity could be used for the FPs which could then borrow the features of the phone (roundedness, position, etc.) from either /@/ or /V/. In this way, the phone identity uniquely identifies the FP vowel, and consonants in the immediate context such as the /m/ in 'UM', while still sharing characteristics of its parent vowel.

7.2.1 Methodology

A preference test was performed to determine whether this alternative representation results in better FP realisation. SiRe (see Chapter 10) was used as the front-end and was modified to convert all vowels in FPs into one of four phones – /V/, /@/, /UHV/ or /UH@/. /UHV/ used the phone features of /V/, and /UH@/ the features of /@/. Four voices, each using one of the four FP phone representations, were trained using HTS-2.3beta (Zen et al., 2007a) and a combined corpus of the read and spontaneous speech (same training set for each voice). This combination was done to ensure a higher overall quality of speech from the read speech corpus while still retaining samples of the FP phone from the spontaneous corpus. Data mixing is discussed in more detail in Section 7.3. 20 sentences containing FPs were selected from a corpus of "found data" derived from transcripts of the BBC's Desert Island Discs (DID) programme and made available as part of the EPSRC-funded *Natural Speech Technology* project. Specifically, the sentences were selected from the utterances of the presenter, Kirsty Young. The upper bound on the length was 25 tokens, the lower 5, and each utterance contained at least one FP. These sentences were synthesised using each of the four voices.

30 paid native English speakers were recruited to take part in the listening test. Each participant rated all 20 sentences for each preference pair. As there are four different voices this results in six pairs of 20 sentences for a total of 120 pairs rated by each participant and a total of 600 ratings of each pair. As the quality of the FPs was of particular interest, and not just the overall quality of the speech, participants were

instructed to “judge which of the two sentences you think sounds the most natural paying particular attention to the realisation of ‘UH’ and ‘UM’ ”. As shown in Chapter 3, naturalness ratings can be influenced by the instructions, and so the above wording was crafted in an attempt to ensure participants focused primarily on the FP realisations. The options were “Sample 1”, “Sample 2” or “No Preference”. Experimental materials can be found at the thesis repository (Dall, 2017).

7.2.2 Results and Discussion

Table 7.1 summarises the results of the preference test. “No Preference” judgements were split evenly over the two systems – this enables the use of the exact binomial test. /@/ was significantly dis-preferred compared to all other representations. There were no statistically significant differences between all other representations, although there was a tendency for the FP-specific phones to be slightly preferred over the /V/, with virtually no difference between the two FP-specific phones.

Although no perceptual preference was found for the FP-specific phones over the standard /V/, an analysis of the predicted acoustics show that both match the acoustic characteristics for FPs better. Table 7.2 shows the mean duration and mean F0 for ‘UH’, ‘UM’ and vowels for each FP phone representation in the synthesised output. These data show that both /UHV/ and /UHV/ have longer durations for FPs (about 100 ms longer) than /V/ and /@/ and that the duration for vowels is roughly equal across all FP models. Furthermore, for both /UHV/ and /UH@/, a lower mean F0 is found. These longer durations and lowered F0 values for /UHV/ and /UH@/ more closely match those found in Chapter 4 (repeated in Table 7.2) and thus show that the FP specific representations better capture the general acoustic characteristics of FPs than /V/ and /@/. Note that while the F0 for the FPs directly compared to the spontaneous speech does not suggest a fall should be exhibited, it is worth remembering that the model here is based on directly combining the read and spontaneous speech which means the natural numbers are perhaps a little misleading – thus the inclusion of the read speech numbers which show how the general model has also been affected by this data combination (more in the next section). The choice between /UHV/ and /UH@/ was made based on the finding that the /V/ phone was significantly preferred over /@/, thus favouring features borrowed from /V/ as it would seem /@/ is not a suitable underlying phone. Therefore, /UHV/ was used in the following investigation.

/V/	/@/	/UHV/	/UH@/	No Preference	Adjusted p
47.2%	35.7%	-	-	17.2%	<0.05
40.0%	-	**43.7%**	-	16.3%	=0.39
39.5%	-	-	**44.3%**	16.2%	=0.25
-	36.0%	**47.8%**	-	16.2%	<0.005
-	34.7%	-	**48.6%**	16.7%	<0.001
-	-	39.2%	**39.5%**	21.3%	=0.97

Table 7.1: Preference test results. P is calculated using the exact binomial test, the preferred phone in a pair is indicated using bold face.

	/V/	/@/	/UHV/	/UH@/	Natural
UH dur (s)	0.152	0.153	0.246	0.250	0.221
UM dur	0.313	0.303	0.431	0.404	0.374
Vowel dur	0.074	0.074	0.079	0.081	0.061/0.068
UH F0 (Hz)	176	175	160	160	175
UM F0	174	177	171	173	170
Vowel F0	169	170	168	169	173/188

Table 7.2: Mean duration and mean F0 for UH, UM, vowels. In all cases there were 30 UH, 9 UM and 717 vowels. The synthesis system used was the straight combination from Section 7.3. For the natural vowel duration and F0 the second number is that for Read speech.

7.3 Data Mixing for FP Synthesis

The improved FP synthesis obtained by Andersson (2013) was achieved by a data-mixing technique previously used for producing various emotions and speaking styles (Yamagishi et al., 2005, 2003). The technique involves training a single model of speech using both read and spontaneous data simultaneously, but distinguishing the two speech types through an added linguistic feature which denotes the speech type the data came from. This affects the decision tree context-clustering, enabling each speech type to be clustered separately during training. At synthesis time each sentence is then marked with either the read or spontaneous tag, so that all speech is steered toward that particular style. The benefit of the method comes from the fact that not all context clusters will be split on the speech-type feature, and therefore some sharing of data is possible. Andersson (2013) concludes that it is that which enables the spontaneous speech voice to match the read speech voice by overcoming data sparsity issues. It was not, however, reported how this method compares to a system that combines the two types of speech in training to produce a voice without marking the speech type. Consequently, that system is here included alongside other methods.

It is possible that the TTS system trained on read speech in Andersson (2013) was matched by the spontaneous speech-based system primarily because it was unable to utilise the FPs present in the spontaneous speech effectively. This seems particularly plausible considering the finding in Section 3.4 that a standard read speech voice is considered more natural than a spontaneous speech-based voice, even when including FPs. This could happen due to the use of the /V/ phone, which would have samples very different from the FPs in the read speech. There are two possible ways to alleviate this. First, we can use an FP-specific phone, and the results in Section 7.2 suggest that /UHV/ seems most promising. This would distinguish the phone from those present in the read data, and therefore, during synthesis, there would be no samples of this particular phone with the read tag. This would force the system to use the spontaneous speech-based /UHV/ model. It may, however, also simply result in the decision tree relying on the features of the phone to utilise the read speech samples of /V/, and so another method of synthesising from the data mixed voice was also applied. In the previous method by Andersson (2013), the sentence to be synthesised was tagged as either entirely read or else entirely spontaneous – but it is also possible to tag only parts of the sentence as coming from either type of speech. Specifically, the proposal here is to tag all of the sentence as read *except* the FPs themselves, which are tagged as spontaneous. This should allow the voice to retain the generally higher overall quality of the read speech, while still synthesising FPs from the more appropriate spontaneous speech model. The main potential problem with this method is that there is no data available of sentences in which this switch happens. However, it seems likely that the trajectory modelling applied should smooth the transitions effectively.

There is also an alternative way of mixing the data, namely by using speaker adaptation. In this approach, an initial model from several speakers is usually trained, before being adapted to a target speaker using adaptation techniques such as the constrained structural maximum a posteriori linear regression (CSMAPLR) technique of Yamagishi et al. (2009). We can apply this technique to the switch between read and spontaneous data by first training a voice on one type of speech and subsequently adapting it to the other. Adapting from read to spontaneous could solve data sparsity issues in a similar manner to the data marking technique, whereas adapting from spontaneous to read could retain read speech quality while still providing data for the FPs in a similar manner to the proposed switch in speech mark when using the marking technique.

7.3.1 Methodology

Four different voices were trained. One was a standard HMM voice in which both the read and spontaneous speech were pooled and used as training data. This provides the baseline approach (and therefore it is the system used in the phone experiment in Section 7.2). Another voice was trained using the data marking technique, and three methods of synthesis were applied: everything marked as read, everything marked as spontaneous, or everything marked as read except the FPs (which were marked as spontaneous). The final two voices were speaker-adaptive voices. One was adapted from a base read model to the spontaneous speech, and the other from spontaneous to read. In total six different synthesis methods were evaluated. The /UHV/ phone representation from the phone experiment was used in all cases, since it showed the most promise and could also potentially help the “everything marked as read” synthesis in realising convincing FPs (as discussed above).

The same 20 sentences from the experiment in Section 7.2 were used, but this time a MUSHRA-style naturalness test was run. This was done in part due to the many preference pairs which would have been necessary, but also in part because we were interested in the overall quality of the resulting speech and not merely the synthesis of the FPs. An additional sentence from the DID corpus was synthesised and used as a training sample. The test was run without a natural reference as the DID data consisted of Kirsty Young's speech, not the voice talent's whose speech was used to base the synthetic systems on. The participants were instructed to rate how natural each sample sounded without further qualification. However, it was explicitly mentioned that conversational phenomena such as FPs would occur, and that this was part of the test. This was done to ensure that participants paid attention to the FPs specifically, while also focusing on the overall naturalness of each voice. Aside from that, the experiment was identical to a standard MUSHRA test with one screen per sentence on which participants listened to and rated all samples of that sentence for each system side by side. This provided both a measure of naturalness and order of preference between the synthetic voices. 30 paid native English speakers were recruited. Each participant sat in a sound-proof booth in front of a computer wearing high quality headphones and rated all 20 sentences for a total of 600 evaluations of each voice. Participants were instructed to rate for general naturalness, however were made explicitly aware that FPs were present in the samples (although not instructed to pay any particular attention to these as done in the previous experiment). The test took approximately 30 minutes to complete for each person. Experimental materials can be found at the thesis repository (Dall, 2017).

7.3.2 Results and Discussion

The results are given in Figure 7.1. All the system pairs were compared using a Wilcoxon signed-rank test, after Holm-Bonferroni correction to avoid false positives. All the systems are significantly different ($p < 0.001$) from each other except for Mark_R (synthesis marked as read) and Mark_Sw (synthesis marked as read but switched to spontaneous for FPs). This finding is somewhat surprising as although it was expected that Mark_Sw would improve the naturalness of the speech, it was not expected that Mark_R would do equally well. A possibility is, that despite everything being marked read in the Mark_R system, spontaneous speech is necessarily used because there are no occurrences of /UHV/ in the read data.

To test this an extra ad-hoc preference test was run to ascertain whether it is the specific phone /UHV/ for FPs that is benefiting Mark_R. 10 listeners took part in this listening test comparing a system with everything marked read using the /V/ phone for FPs versus a system using the /UHV/ phone for FPs. Each subject rated 45 sentences,

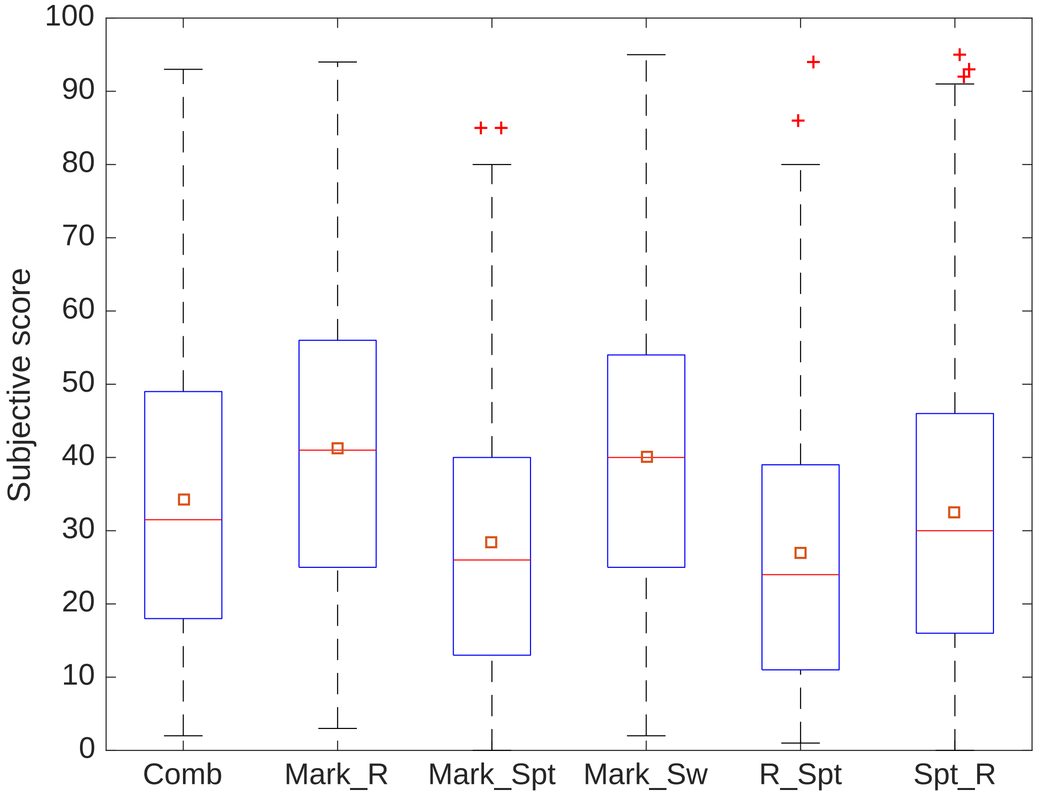


Figure 7.1: Results of the MUSHRA-style test. Comb = Direct combination with no source marking. Mark_R = Combination with source marking synthesising with all marked as read, Mark_Spt = Combination with source marking synthesising with all marked as spontaneous, Mark_Sw = Combination with source marking synthesising with FPs marked a spontaneous and the rest as read. R_Spt = read adapted to spontaneous. Spt_R = Spontaneous adapted to read. Red line is the median, square the mean.

	Read /V/	Read /UHV/	Spont /UHV/	Switch /UHV/	Natural
UH dur (s)	0.059	0.143	0.241	0.242	0.221
UM dur	0.134	0.303	0.416	0.324	0.374
Vowel dur	0.078	0.080	0.075	0.086	0.061/0.068
UH F0 (Hz)	182	183	171	166	175
UM F0	183	183	166	167	170
Vowel F0	185	186	166	184	173/188

Table 7.3: Mean duration and mean F0 for UH UM and vowels. In all cases there were 30 UH, 9 UM and 717 vowels. Synthesis systems used: Mark_R (V), Mark_R (UHV), Mark_Spt, Mark_Sw.

an extended set of DID materials. Instructions and listening conditions were as detailed in Section 7.2. In this test, listeners preferred the combination system with /UHV/ phone 63% of the time, significant using the exact binomial test ($p < 0.0001$). This suggests that the use of the /UHV/ phone did indeed allow the read marked speech to utilise the spontaneous FP occurrences to inform its model and that this improved the overall perception of the voice.

Furthermore, if we compare the FP durations and F0 for the marked system using /UHV/ when either marking all as read, spontaneous, or switching with the read marked system using /V/ (Table 7.3) we can see that the read marked /UHV/ system produces durations and F0 closer to those expected for spontaneous speech, whereas the read marked system using /V/ does not capture this at all. However, using the mark switching technique gave even better results, particularly when compared to the natural speech. This suggests that although no perceptual preference difference was found between the read marked and switching system using /UHV/, the switching system still better captures the acoustic realisation of FPs.

Interestingly, neither adaptation system performed very well. Informal listening suggests that despite the two types of speech being from the same speaker, adaptation still introduces many serious artefacts which degrade the overall speech quality.

Data mixing using a mark switching technique and a specific phonetic representation thus seems like the most successful approach to FP modelling, both from a perceptual and acoustic standpoint. However, there is one potential limitation to the approach. The switch between the reading mode of the read speech to the speaking mode of the spontaneous at the point of FP synthesis can cause a discontinuity in the prosodic realisation of the speech which may affect the llistener. While the approach does indeed result in perceptually preferred speech, this discontinuity could lead to a subconscious degradation of the speech understanding, and thus affect e.g. reaction time. In Chapter 10 it will be investigated whether a voice utilising this switching technique can reproduce the reaction time experiment findings from Chapter 4.

7.4 Chapter Discussion and Conclusions

This chapter has focused on the topic of modelling FPs overtly in TTS. Initially a test comparing read and spontaneous speech-based voices using FPs was presented. The spontaneous speech-based voice did not match the read speech-based voice despite the presence of FPs. This contrasted with earlier findings that such voices could match voices trained on standard corpora (Adell et al., 2012; Andersson, 2013) when synthesising sentences containing FPs. However, in both cases the systems used were modified forms of a standard TTS system.

Furthermore, the FPs produced by both read and spontaneous speech-based voices did not exhibit the acoustic characteristics that have been shown in the literature Adell et al. (2012); Shriberg (2001); O'Shaughnessy (1992) and in the investigation of the CS corpus presented in Chapter 4 – notably longer durations and lower F0 compared to fluent speech.

Consequently, a number of different phonetic representations for modelling FPs were investigated, this was to find whether having a distinct phone for FPs would capture the acoustic properties of FPs more successfully (similar to the modelling of Adell et al. (2012)). On the basis of a preference test and acoustic analysis, the FP specific phone /UHV/ was deemed the best for FP modelling. It was significantly preferred over /@/ and the longer durations and lower F0 more closely match the desired acoustic characteristics than /V/.

In addition to a specific phone for FPs, various data-mixing approaches to using both the read and spontaneous speech were explored - straightforward combination of read and spontaneous data, data source marking and speaker adaptation. It was found that a data-marking technique similar to Andersson et al. (2012) performed the best. However, in contrast to Andersson et al. (2012), this technique did not improve the spontaneous speech-based voice to match that of a read speech-based voice. The results obtained here used a specific FP phone representation, and a preference test showed that this representation improved the perceived synthesis quality. This suggests that the voice based on read speech, as in Andersson et al. (2012), suffered degrading quality issues due to the bad FP representation of /V/, as using the /V/ phone tended to use read data in which no FPs occurred.

By using a specific FP phone /UHV/ it was here found that a data-mixed voice synthesising with a read speech tag could produce more convincing FPs such that perceptual quality did not degrade compared to a voice in which the FPs were synthesised using the spontaneous mark. However, using the spontaneous mark produced FPs even closer to the expected acoustic properties, and thus the switching of the mark, from read in general, to spontaneous when synthesising FPs, produced perceptually high quality speech while also retaining the defining characteristics of FPs.

Chapter 8

Alternative Linguistic Context Feature Sets

A Note About Collaborative Work

The parsing-based linguistic features and the positional representation work was carried out while at Nagoya Institute of Technology and has been published, in regards to read speech only, as Dall et al. (2016b). Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku and Keiichi Tokuda (all at Nagoya Institute of Technology) were all invaluable discussion partners for this work and the set of features to extract was agreed upon in collaboration. Special mention goes to Kei Hashimoto for his help in setting up and using the DNN system. They were not involved in the extension of this work which incorporated the datamixed voice.

8.1 Introduction

In the previous chapter, we saw how explicit modelling and data mixing could improve the synthesis of filled pauses. In this chapter, we turn to another method of improving the overall quality of synthetic speech, namely incorporating linguistic features based on parsing.

This falls under the category of conversational TTS due to its potential to identify FPs as being in a special phrase structure and thus different from other words/text tokens, but also because of its potential to improve the phrase level prosody of TTS – particularly important in conversation.

In HMM synthesis, a linguistic context feature set is used for decision tree based clustering of the training data (see Chapter 2.3.2). The set of contexts used is derived from linguistic analysis of the text of the training corpus, and contains information about the context of the current phone, syllable, word, phrase and utterance. The standard set of features used in HMM synthesis was proposed together with the first release of the HTS speech synthesis system (Zen, 2002), and notably has not changed since then (HTS, 2015) (see Section 8.3 for details of this set). This is not to say that researchers have not been interested in this area, however, there is little direct research on the topic.

Additionally, research has shown that, in the current feature set, features above the word level have little to no impact on final speech output quality. In Watts et al. (2010) features were slowly removed starting from utterance level features down to syllable level, and it was found that no above word level features had any significant impact on the resulting models. In a similar vein, Lu and King (2012) used a Bayesian network to find the most relevant features from the standard set, and obtained a minimal degradation of output speech quality with a feature set consisting of only 6-9 (depending on stream) features, down from the standard 26. This suggests that many features at the word-level and above do not have a large impact on speech quality. However, it should be possible to find such features.

In particular, the use of parsing derived features has been investigated in a few studies. The system proposed by Yu et al. (2013a,b) incorporates parsing derived features, but unfortunately they do not specify which type of parsing (likely Probabilistic Context Free Grammar (PCFG)) nor exactly which features were included (Section 3.1.2 in Yu et al. (2013b) and Section 3 in Yu et al. (2013a)). The effect of parsing was also only indirectly evaluated using objective measurements in Chinese (Yu et al., 2013a) and by submitting a, very well-performing, English system for the 2013 Blizzard challenge with similar features (Yu et al., 2013b). It is however unclear whether the system performed well due to the addition of the parsing features or due to other differences, e.g., that it was a hybrid system.[1] In Suni and Vainio (2008), a Finnish system for using parsing derived features for rule-based prominence prediction is described, however they do not put the system to the test. In the Spanish system from Barra-Chicote et al. (2010) (Section 4.3), morphosyntactic features derived from part-of-speech tagging and a parse tree were used. These features improved the synthetic speech quality, but unfortunately which features were used was not detailed. In French, Obin et al.

[1] A hybrid system is generally a unit selection system in which the unit selection is driven by a statistical model.

(2010) use the Alpage Linguistic Processing Chain to extract a set of parsing derived features and, in a comparison mean opinion score (CMOS), certain types of sentences show an improvement, while others degrade. Their baseline did, however, include some morphosyntactic features, which are, at least in English, not standard and are arguably parsing derived. Overall this suggests that parsing based features can capture information relevant to TTS systems at the word-level and above.

Furthermore, deep neural network (DNN) synthesis is becoming increasingly popular and this method of synthesis often combines the decision tree context clustering and acoustic modelling in the neural network (e.g., Zen et al. (2013); Wu et al. (2015)). Because of this it is likely that the neural network may be able to utilise different features than the HMM system. As such, this work will also focus on some alternative representations of positional values. Positional values are features related to, e.g., phoneme position in syllable or word position in utterance. These values are currently represented as forwards and backwards absolute positional values. Absolute values, however, introduce ambiguity about the meaning of features in segments of differing length, so, in Section 8.3, an investigation into the use of relational or categorical values as an alternative is presented. Note that this investigation is also novel with regards to HMMs, at least this author has not found any papers suggesting it has been investigated.

Although this thesis is generally not concerned with DNN synthesis, it is included in this particular chapter. This is primarily due to the fact that the mainstay of the work was done while the thesis author was visiting Nagoya Institute of Technology – during which the focus was mainly on read speech corpus quality. Although the DNN results could have been excluded it was felt that having them here for completeness was better as some choices were based on the DNN results in combination with the HMM results – and not including them would thus mask those choices. DNN synthesis is not included in the final extension of a voice using both read and spontaneous speech for consistency with the rest of the thesis.

After the investigation into positional features, two sets of additional features derived from PCFG and dependency parsing are presented. The choice of PCFG and dependency parsing is in part due to their likely use in the above mentioned studies, and also due to them being the perhaps two most popular types of parsing used within the Natural Language Processing community. Indeed both of these are methods with a long history and have well defined toolkits, e.g., the Berkeley (Petrov and Klein, 2007) and Stanford (Klein and Manning, 2003) parsers, standard algorithms and active research communities (e.g., Andreas et al. (2015); Chen and Manning (2014); Durrett and Klein (2015); Socher et al. (2013)). In Section 8.4, each method of parsing is presented, a derived set of additional features from each parsing method is also presented followed by an evaluation of the effect of using these features in HMM and DNN synthesis. Section 8.5 will provide an overall discussion of the findings for read speech based voices. Finally, in Section 8.6, an evaluation is presented of the resulting system when training a datamixed voice of the kind presented in the previous chapter is used to synthesise sentences containing FPs.

The resulting system is not evaluated on spontaneous data alone, but only in the datamixed case. There are two reasons for this. Firstly, transcriptions of spontaneous speech are notoriously difficult for parsers to deal with, so parses derived from this type of speech corpus are more likely to be incorrect and thus likely to be less useful. Secondly, as has been shown several times in this thesis, voices based on spontaneous speech alone are currently simply not as good as their read speech counterparts – even after several improvements – and the improvements gained from applying parsing is unlikely to bridge this gap. However, the evaluation of a datamixed voice is reasonable as the parsers *are* good at identifying FPs and could potentially further improve their modelling while also benefitting from the improved read speech voices attained from parsing.

8.2 Overview of Current Linguistic Features

The linguistic feature context set for HMM-based speech synthesis has remained near constant since the first release of HTS. Comparing the default set of the newest HTS release (2.3) and that proposed in the first HTS release (1.0) we can note that they are, in fact, identical[2]. Furthermore, if we compare the feature sets produced by a number of the most well-known and freely available front-ends, see Table 8.1, we see very little variety. Festival (Black et al., 2014) and Flite (Black and Lenzo, 2001) are exactly equivalent (on purpose), and Idlak (Aylett et al., 2014) uses a subset of the Festival/Flite set, while MaryTTS (Schröder and Trouvain, 2003) is overall very similar to Festival/Flite.[3]

[2] Both releases can be found at the HTS website.

[3] The set of features presented for MaryTTS was derived through conversation with Sbastien Le Maguer, one of the current maintainers of the MaryTTS system.

	HTS 2.3	HTS 1.0	Festival / Flite	MaryTTS	Idlak
Quinphoneme Context	✓	✓	✓	✓	✓
Phoneme Syllable Position (Forward, Backward)	✓	✓	✓	✓	✓
Phoneme Word Position (Forward, Backward)	x	x	x	✓	x
(Previous, Next) Phoneme Pause	x	x	x	✓	x
Phoneme at Utterance (End, Beginning)	x	x	x	✓	x
Trisyllable Stress	✓	✓	✓	✓	✓
Trisyllable Accent	✓	✓	✓	✓	x
Trisyllable Number of Phonemes	✓	✓	✓	✓	✓
Current Syllable (Phrase, Word) Position (Forward, Backward)	✓	✓	✓	✓	x
Number (Stressed, Accented) Syllables (Before, After) Current Syllable in Phrase	✓	✓	✓	✓	x
Number of Syllables (To, From) Next (Stressed, Accented) Syllable	✓	✓	✓	x	x
Current Syllable Vowel ID	✓	✓	✓	x	x
(Current, Next, Next-Next) Syllable ToBI Accent	x	x	x	✓	x
(Current, Next, Next-Next) Syllable ToBI End-Tone	x	x	x	✓	x
(Current, Previous) Syllable Break	x	x	x	✓	x
Triword General Part-Of-Speech	✓	✓	✓	x	x
Triword Part-Of-Speech	x	x	x	x	✓

Triword Number of Phonemes	x	x	x	✓	x
Triword Number of Syllables	✓	✓	✓	✓	✓
Triword Punctuation?	x	x	x	✓	x
Number of Words (To, From) Punctuation	x	x	x	✓	x
Current Word Phrase Position (Forward, Backward)	✓	✓	✓	✓	x
Current Word Utterance Position (Forward, Backward)	x	x	x	✓	x
Content Words (Before, After) Current Word in Phrase	✓	✓	✓	x	x
Number Words (To, From) Content Word in Phrase	✓	✓	✓	x	x
Triphrase Number of (Syllables, Words)	✓	✓	✓	x	x
Current Phrase Number of Words	✓	✓	✓	✓	✓
Phrase ToBI Endtone	✓	✓	✓	✓	✓
Previous Phrase ToBI Endtone	x	x	x	✓	x
Phrase Utterance Position (Forward, Backward)	✓	✓	✓	✓	x
Utterance Number of Syllables	✓	✓	✓	x	x
Utterance Number of (Words, Phrases)	✓	✓	✓	✓	x

Table 8.1: A comparison of the feature sets of popular front-ends. HTS standard set uses Festival as the front-end and is therefore equivalent but presented for completeness.

The most in-depth linguistic analysis comes in the form of Part-of-Speech (POS) tagging (often reduced to a few more general tags as in Festival/Flite), accent prediction (last stressed syllable of a content word) and sometimes ToBI (Silverman et al., 1992) labelling. With regards to ToBI labelling only MaryTTS uses anything but end-tone labelling, and end-tone labelling is often just whether the phrase ends in a question mark or not, as in Idlak (Aylett et al., 2014).

For this work, the Festival/Flite set will be treated as the base set, which is also the set used by HTS as Festival is normally used as the front-end.

8.3 Redefining Positional Values

As can be seen in Table 8.1, many of the linguistic features are positional values. These are expressed as the forward and backward absolute position of a segment in a larger segment, e.g., the position of a word in the utterance. So the word "hit" in:

$$The\ man\ hit\ the\ brown\ dog. \tag{8.1}$$

has a forward positional value of 3, as it is the third word from the beginning, and a backward position of 4, as it is fourth from the end. There is however an issue with this representation as the values can quickly become seemingly meaningless. If the sentence was "Peter hit the dog" or "the angry man hit the dog" both positional values change without seemingly affecting the word's pronunciation much. This issue is much more pronounced at the beginning or end of a segment as the meaning of, e.g., a forward position of 5 is very dependent on the length of the segment, if the segment is 5 long it is at the end, but if it is 10 long it is in the middle.

The proposal here is therefore two alternative representations which may alleviate this issue. These alternatives are to represent the positional values as either relational or categorical positions. A relational position is the position of the shorter segment in the longer, normalised to a value between 0 and 1 where 0 is the beginning and 1 the end. This removes the need for forward and backward positioning as the relational forward position of 0.1 is always equivalent to the backward of 0.9, reducing the number of features necessary while capturing the same information. But, it does not entirely solve the issue; as positional values are now normalised you get small differences between positions with every small change in segment length, e.g., in Sentence (8.1) the relational position of "hit" would change from 0.4 to 0.5 if it was changed to "the man hit the dog". This necessitates many additional questions in the questions set, particularly greater/less than questions, due to the increased number of possible values in use as the relational position changes when segment length changes. This shouldn't be an issue for the DNN system as it uses continuous input values naturally, though a decision tree for HMM context clustering may have problems modelling this.

Another way of representing the positional values is to use categories. The proposal here is to use the following 4 categories; "beginning" for the first element, "end" for the last, "one" for segments of length one and "middle" for all others. This reduces the number of potential context values to a great extent, however, it results in a loss of granularity. In order to retain some context for segments close to the edges, the positional category of the previous and following segment is also added.

Thus we have three ways of representing positional values:

- Absolute: The absolute forwards and backwards position of a segment in a larger segment (e.g. word in phrase).
- Relational: The relational position of a segment in a larger segment as a value between 0 and 1 (0 is beginning, 1 the end, 0.5 the middle).
- Categorical: The categorical position of a segment in a larger segment and it's previous and next segment value ("beginning", "end", "one", "middle").

Table 8.2 shows the size of the question set for the HMM, the size of the input layer for the DNN and the total number of decision tree leaves for the different representations after voice training (see next section for HMM and DNN system details). The question set size is after pruning questions not relevant for the training dataset for each of the three methods. As can be seen, the categorical representation results in a decrease in the number of questions, whereas the relational increases it. For the DNN, the size of the input layer increases when using the categorical representation. This is due to the addition of previous and following contexts for edges which necessitates adding additional nodes to the layer.

While the number of questions does not indicate the quality of a feature set, if we look at the number of leaves of the learned decision trees (which can be seen as an indication of quality due to it showing the tree has managed to find more meaningful clusters on the same data) in Table 8.2, all three representations result in approximately similar sized trees despite different question set sizes. This suggests that the categorical representation can, with fewer questions, capture the same information as the other two representations, but, by the same token it seems the relational representation needs

	Positional Representation Type		
	Absolute	**Relational**	**Categorical**
HMM Question Set Size	1264	1714	1059
DNN Input Layer Size	510	503	561
HMM Decision Tree Leaves	6015	5999	5937

Table 8.2: The size of the HMM question set when considering only questions pertaining to the data, DNN input layer size when training a voice and F0 decision tree leaves of trained voice for different positional representations.

more questions. However, as all three approaches result in similar sized decision trees this could be an indication that different sets do not result in different synthesis – but this is not necessarily true – so to test this a small listening test was performed.

8.3.1 Evaluating Positional Values

In order to evaluate the three methods of representation, HMM and DNN voices were trained using each of them. The full corpus of read speech (1974 sentences, approximately 2 hours) from Chapter 3 was used. The HMM systems were built using HTS 2.3 Beta. The DNN systems were built using a system similar to the DNN-MLPG system of Hashimoto et al. (2015), but modified to use STRAIGHT for vocoding, and consisted of a 5-layer feed-forward network. The HMM system employs GV and MLSP postfiltering but the DNN does not. However, as the interesting variable is the effect of the feature types in each system separately, this is not an issue, and consequently no comparisons across system type were made.

Each of the three methods of representation was compared within each synthesis method (DNN or HMM) using a preference test. 10 native English speakers were recruited, and they performed the test in sound-proofed booths wearing high-quality headphones. Each participant evaluated each pair of representations, for both HMM and DNN systems, i.e., 6 preference pairs. The same 15 sentences were used for both HMM and DNN systems. Sentences were presented in a random order and the order within a trial was randomised to avoid bias effects. In total, each listener gave their preference for 90 trials (15 x 3 HMM and 15 x 3 DNN) and had the options to pick either sample or “No Preference”. Which means participants were presented with each pair twice – once for the HMM and once for the DNN system – this is acceptable as we are not comparing across the DNN and HMM systems. The test took approximately

	Absolute	**Relational**	**Categorical**	**No Preference**
HMM	42.7%	**47.3%**	-	10.0%
	-	35.3%	**49.3%**	15.3%
	38.0%	-	**46.7%**	15.3%
DNN	**42.7%**	32.0%	-	25.3%
	-	40.0%	**42.0%**	18%
	32.0%	-	**48.0%**	20.0%
Average	38.9%	38.7%	**46.5%**	17.3%

Table 8.3: Preference scores for each representation pair and the total combined preference for the different positional representations. Bold-faced scores are the highest for that pair.

20 minutes to complete. All experimental samples are available at the thesis repository Dall (2017).

All "No Preference" ratings were evenly split between each pair to allow to allow for statistical significance testing. From Table 8.3 we can see that, in all cases, the categorical representation is slightly preferred, whereas the relational is preferred over the absolute in the HMM case and the absolute over the relational in the DNN case. These are not statistically significant on their own, partially due to the small test size, but if we pool all datapoints from HMM and DNN (bottom Table 8.3) we can see that the categorical representation is preferred over the others, and this difference is significant, using the exact binomial test, compared to the absolute and the relational ($p < 0.05$). Although this is a non-standard method, it still highlights the trend which is apparent, that the Categorical position is generally preferred. Therefore, for the rest of this chapter, the categorical representation will be used.

8.4 Parsing Based Features

Two types of parsing were considered as mentioned earlier: probabilistic context free grammar (PCFG) and dependency parsing.

PCFG is a probabilistic form of a syntactic tree-based grammar. A parse tree is generated from a set of rules to create a tree expanding from a root node downward from very general phrase types to part-of-speech tags at the leaf nodes. Each rule expansion has a probability assigned to it, and each leaf a probability of expanding into

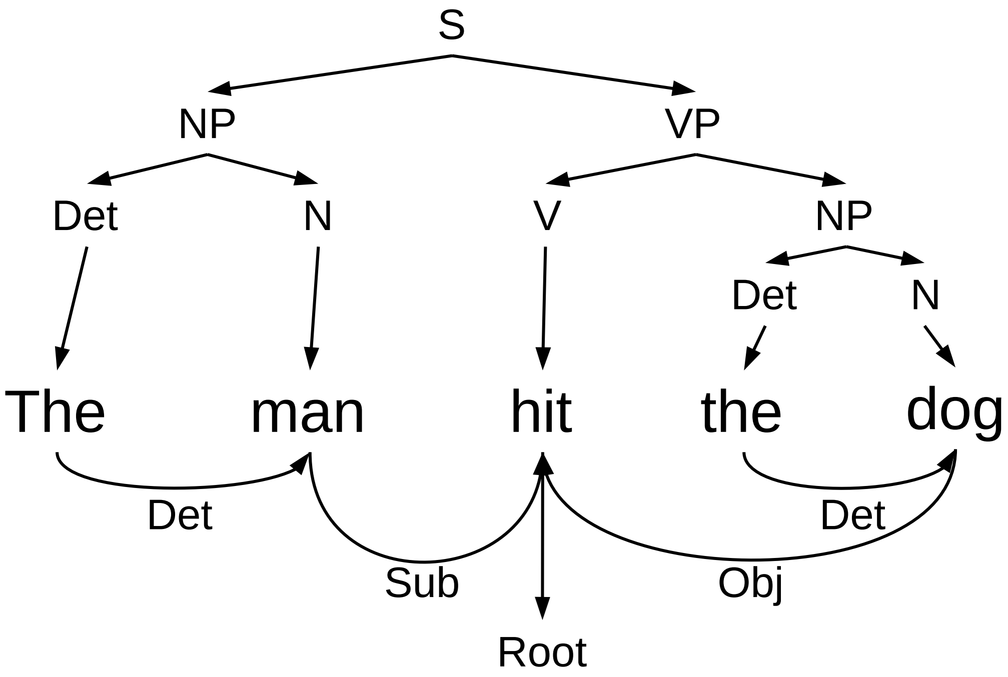


Figure 8.1: A PCFG (top) and dependency (bottom) parse example.

a given word. Finding a PCFG parse involves finding the most likely parse tree given the words and the possible rule-derived trees expanding over those words. A PCFG parse thus describes the syntactic phrase structure of a sentence, something which is likely to be helpful for the overall phrase level prosody – assuming syntax and prosody are related.

A dependency parse, closely related to shallow semantic parsing, describes the internal relations between words in a sentence. These are relations starting from the root of the sentence, the verb, and relations are then found to the rest of the words in the sentence. The relations describe the internal dependencies between words, such that, e.g., the object of the sentence will have a dependency from it to the verb and so will the subject to the verb and so on, until every word stands in exactly one dependency relation to another.

Figure 8.1 illustrates a PCFG and dependency tree for the sentence "The man hit the dog". Note that not all dependency grammars describe a tree structure with exactly one dependency relation upwards for each word, but for the purposes of this work a purely tree-based representation has been used to derive features.

From both of these parses a number of features were extracted. From the PCFG parse a set of features, likely similar to that of Yu et al. (2013b), was extracted which were:

- Great-grand-/grand-/father phrase of the current word
- Position of current/next/previous word in great-grand-/grand-/father phrase
- Expanded general POS-tag

Father phrase is the phrase type associated with the word one node up in the tree, grandfather is the father of the father and so on. If there is no node further up the tree (i.e., we are at the root) an out-of-scope marker ('xx') was used. For the position, the categorical position of the current/previous/next in all of its father etc. phrases was used. The expanded general POS-tag is an expanded version of the general part-of-speech tag category from Festival, which splits the "content words" in the Festival set to slightly more detailed categories such as "verb"/"noun"/"adjective" – but still not the full Penn Treebank (Marcus et al., 1993) set as that was considered too specific.

From the dependency parse the following features were extracted:

- Current word/father/grandfather to father/grandfather/ great-grandfather relation
- Number of children relations
- Tree arc distance to previous/next word
- Current word distance to father/grandfather/great-grandfather word
- General relation to father word

The father etc. relation of a word to another is the relation one up in the parse tree as before. As each word is not at a node in the tree, the number of children relations was also included. The current word distance to father etc. word is the absolute distance in the actual sentence in terms of number of words. The general relation category is defined using the Stanford parser's documentation on the dependency parser (pp. 11-12 of de Marneffe and Manning (2015)) by making most relations one less specific category.

As the features from both types of parsing can be complementary, a combined set was also created, using all features from both parsing methods. This yielded four different sets of features, a standard set equivalent to the current Festival/HTS features (Standard), a set with the PCFG features added (PCFG), one with the dependency features added (Dependency) and a combined set (Combined). PCFG and dependency parses were both extracted using the lexicalized parser of the Stanford Parser version 3.5.1.

8.4.1 Evaluating the Parsing Based Features

For each of the four different sets (Combined, PCFG, Dependency and Standard), an HMM and DNN voice was trained using the same systems as in Section 8.3.1 and using the categorical positional representation. Each of the 4 voices from each type of system was compared to each other in a preference test, this resulted in 12 (6*2) system pairs. 30 native English speakers were recruited and each participant rated 15 sentences for each pair resulting in a total of 180 comparisons per participant. The options were 'Sentence 1', 'Sentence 2' or 'No Preference'. The test had 4 sections of 45 pairs and between each section participants were asked to get out of their booth and walk around a bit before continuing the next section. This was done to avoid listener fatigue. Section order, sentence presentation and pair order was randomised to avoid bias effects. All experimental samples are available at the thesis repository (Dall, 2017).

Table 8.4 summarises the results; all 'No Preference' scores were split evenly between each pair to allow for statistical significance testing. In all cases, the parsed versions are preferred over the unparsed. This difference is significant, using the exact binomial test, when using the Combined set for the HMM ($p < 0.01$) but not when using the PCFG or Dependency sets alone. For the DNN it was significant for the Combined ($p < 0.001$) and PCFG sets ($p < 0.001$) but only marginally for the Dependency set ($p = 0.066$). For the HMM there is very little difference between the three parsed versions, however, for the DNN the Dependency parse is significantly dispreferred (Combined: $p < 0.05$, PCFG: $p < 0.001$) to the other two (while still being prefferred over no parsing), between Combined and PCFG the difference is not significant ($p = 0.54$). As listeners were asked which of the two sentences in each trial they considered most natural this means that the standard unparsed feature set leads to synthetic speech which is considered less natural than speech generated using the parsed sets, and particularly the DNN system was able to take advantage of these new features.

8.5 Discussion of Effect on Read Speech

The current linguistic context set uses absolute positional values. In this investigation, using categorical positional values provided a better representation resulting in synthesis output which was preferred over synthesis using either absolute or relational

	Standard	**PCFG**	**Dependency**	**Combined**	**No Preference**
HMM	45.1%	**48.2%**	-	-	6.7%
	44.0%	-	**48.9%**	-	7.1%
	40.7%	-	-	**53.3%**	6.0%
	-	**47.3%**	44.9%	-	7.8%
	-	43.1%	-	**46.0%**	10.9%
	-	-	45.6%	**46.0%**	8.4%
DNN	36.7%	**52.7%**	-	-	10.7%
	39.3%	-	**48.2%**	-	12.4%
	35.1%	-	-	**53.8%**	11.1%
	-	**59.6%**	33.3%	-	7.1%
	-	**47.1%**	-	43.1%	9.8%
	-	-	39.8%	**50.2%**	10.0%

Table 8.4: Preference scores for each representation pair and the total combined preference for the differing parsing features.

values. This difference was more pronounced for the HMM than the DNN system, but the tendency was the same for both (Table 8.3). That this had less of an effect in the DNN may be due to the way the categorical feature was represented. For the HMM, using the categorical representation reduced the question set, and thus the possibility for mistakes, by around 20%, and this is presumably one of the reasons for the improvement as the decision tree still captures the salient information using fewer features. However, for the DNN the relational and absolute values were normalised and represented by a continuous value using one input node; but the categorical values were represented using several binary nodes, thus increasing the size of the input layer in a way which may have made it harder for the DNN to deal with. It is also possible that the DNN may simply be able to compensate for the more confusing input of the relational and absolute values in other ways than the decision tree for the HMM due to its non-linear nature.

Deriving features based on PCFG and dependency parsing also improved synthesis. This improvement seems primarily to have come from the prosodic domain, with the F0 decision tree clustering being the most affected of the different streams (as shown in Table 8.5). Figure 8.2 shows a sample sentence and the generated F0 contours, from which we can see that all of the parsing feature sets generate a more lively F0 output.

	Parsing Feature Condition			
	Standard	**PCFG**	**Dependence**	**Combined**
HMM Question Set Size	1059	1225	1229	1390
DNN Input Layer Size	561	691	688	817
HMM F0 Decision Tree Leaves	4311	4427	4499	4494

Table 8.5: The size of the HMM question set when considering only questions pertaining to the data, DNN input layer size when training a voice and F0 decision tree leaves of trained voice for each set of parsing features using the categorical positional representation. Only the F0 decision tree is considered as this is where most differences are noted.

This finding makes good sense: the PCFG provides a good representation of the detailed phrase structure of a sentence, improving phrase level F0 movements, and the dependency parse highlights important words in particular relations, improving prominence patterns of a sentence. That these features are complementary is not entirely obvious as the PCFG and Combined sets are equally preferred, it is, however, likely that this may be due to the larger number of features in the Combined set.

It has been argued that one potential reason why the categorical representation performs better than the others is that it describes the same information using fewer features, however, many additional features have then been added through the parsing derived sets. Here no attempt has been made to weed out features that are not useful, and consequently the size of the feature set has increased – which could lead to increased levels of confusion in the models. Applying methods for feature set reduction, such as in Lu and King (2012), could help reduce the feature set without affecting the performance, in fact it may help the larger Combined set. However, it does appear that the DNN, which saw the clearest improvement and the largest increase of input size, is capable of dealing with these additional features to some extent. Furthermore, by using the categorical representation, the total number of questions for the Combined set (Table 8.5) is only a small increase compared to the standard set using absolute positional values (Table 8.2) for the HMM. Applying feature reduction methods would be a useful excersise, however, is considered out of scope in this thesis.

There are many different potential features to derive from parsing, the sets presented here are by no means the only combinations, they do not use the only possible parsers and are unlikely to be the best combinations. However, these extensions and

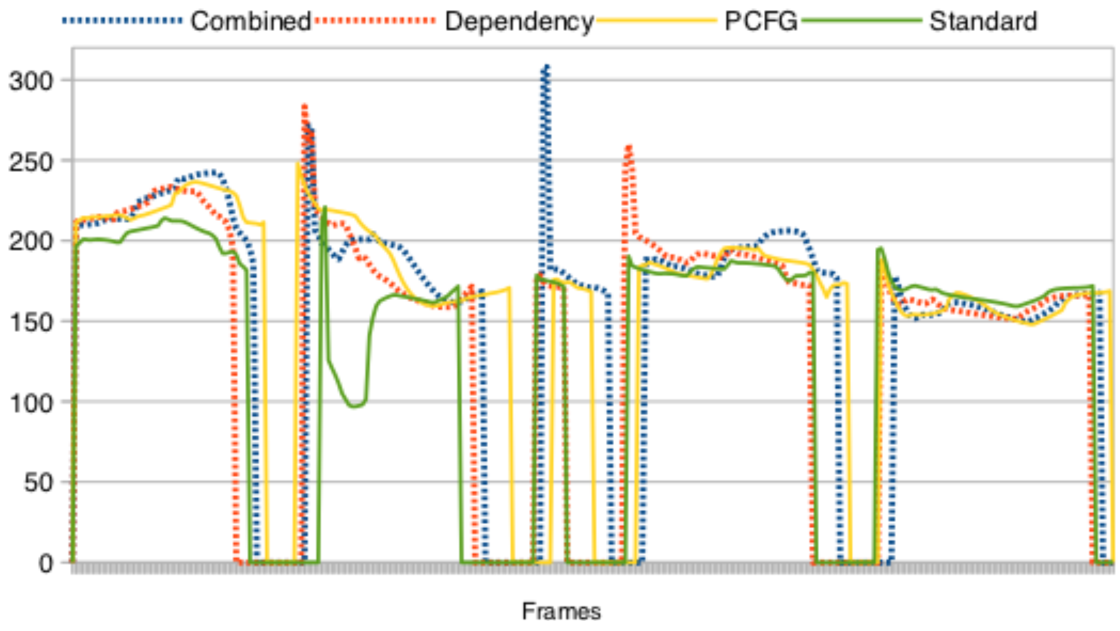


Figure 8.2: Generated F0 contours for a sample sentence.

investigations are also beyond the scope of this thesis, it is hoped however, that others may wish to refine these sets – or find new ones – and to this end a research front-end written in Python 2.x which can reproduce all of the presented context sets, and which allows for rapid experimentation with alternative sets is released as part of this thesis. Details are presented in Section 10.4.

8.6 Applying Parsing Derived Features to Filled Pause Synthesis

It could be claimed that using parsing derived features to improve TTS voices using read speech may be good, but not directly relevant to the main purposes of this thesis, which is the utilisation of spontaneous speech and the introduction of conversational elements into TTS voices. Arguably, however, that the improvement is primarily in the prosodic domain is of relevance, as richer prosody is very much an element of conversational speech. The use of only read speech less so. One reason for not including spontaneous speech initially is because of the questionable ability of parsers (Bechet et al., 2014; Zechner and Waibel, 1998) to deal with sentences which are not well-formed and thus initially verifying the approach on read speech is useful. Now that it has been shown that read speech based voices benefit from parsing, we can turn toward utilising spontaneous speech.

Standard	Categorical+Parsing	No Preference
35.9%	51.9%	13.7%

Table 8.6: Preference scores between the disfluent voices with the standard features or the proposed features.

Specifically, the parsing will not be applied only to spontaneous speech, for the above mentioned reason of the parsers (in)ability to deal well with this type of speech, but also because of the inability of current TTS methods to model spontaneous speech well – something this is unlikely to rectify. But, parsing *is* good for identifying filled pauses, and thus through the parse it can be made obvious that the FPs are in a special utterance structure and should help their modelling and the modelling of surrounding phones.

8.6.1 Preference Test

A datamixed HMM voice of the type described in Chapter 7 was trained using the standard linguistic feature set and one using categorical positional values and the Combined set of parsing derived features – as this was the feature set found to yield the highest preference scores. It utilised the same 1k read and 1k spontaneous speech corpora as the rest of the thesis. FPs were represented by the /UHV/ phone found to be the best in the previous chapter.

A preference test was run, again measuring naturalness, for which 30 paid native English speakers were recruited. The same set of 45 FP containing sentences from Section 7.3.2 was used. Participants were instructed to rate according to general naturalness, without further qualification. Each participant rated each pair of sentences once and the choices were "Sample 1", "Sample 2" and "No Preference". The order of presentation of the pairs was randomised for each participant, and so was the order of presentation of the utterances within each pair. In total this yielded 1350 ratings. Experimental materials can be found at the thesis repository (Dall, 2017).

Table 8.6 shows the preference scores between the pairs. "No Preference" ratings were evenly divided between the systems to allow for significance testing using the exact binomial test. The proposed system is significantly ($p<0.001$) preferred over the standard system, showing that the categorical and parsing based features combined yield improvements even when spontaneous speech and filled pauses are included.

If we look at the duration and F0 of the filled pauses produced (Table 8.7) we

	Parsed	Unparsed	Natural
UH dur (s)	0.211	0.241	0.221
UM dur	0.265	0.381	0.373
Vowel dur	0.083	0.084	0.061/0.068
UH F0 (Hz)	161	148	175
UM F0	174	186	170
Vowel F0	184	187	173/188

Table 8.7: Mean duration and mean F0 for 'UH', 'UM', vowels. In all cases there were 30 UH, 25 UM and 959 vowels. For the natural speech vowel durations the first number is the spontaneous speech and the second the read.

can see that both the parsed and unparsed versions exhibit the general tendencies of increased durations and lowered F0. They do, however, do it in a slightly differing manner: the parsed set produces generally shorter durations and lower F0 for 'UM', and the unparsed longer durations and lower F0 for 'UH', but with no effect for 'UM'. It is thus hard to say which best represents the natural effect, and thus the better FP representation. The preference effect is thus likely to be due to the parsing and/or the positional value representation.

So, despite the difficulty of parsers to correctly parse the spontaneous speech, applying parsing and categorical positional values still improve preference for the parsed set in a datamixed voice utilising spontaneous speech for FP synthesis. It is possible that the inclusion of filled pauses helped the parsers get the correct parse, as these are more reliably identified, and, although parsing did not help FP synthesis specifically, thus could have provided reasonable parses to derive features from.

8.7 Conclusions

In this chapter, an investigation of the linguistic context feature set for HMM and DNN synthesis has been presented. A large part of the standard context feature set utilises positional values for a segment in a larger segment. It has been shown that a categorical representation is preferred over the standard absolute and another alternative representation using relative values. Additionally, the use of PCFG and dependency parsing can provide additional features, useful for describing the word level interactions, particularly with regard to F0, and that using these features for voice building improves the resulting synthesis. These improvements carry over to the use of a datamixed voice including spontaneous speech, despite spontaneous speech being less readily parsable.

Chapter 9

Automatic Insertion of Filled Pauses Into Text

9.1 A Note About Collaborative Work

The work in this chapter was done in close collaboration with Marcus Tomalin (University of Cambridge) and Mirjam Wester (CSTR). In particular, Marcus did most of the work related to the SVM and Decision Tree methods, the expansion to use multiple disfluency types and the disfluency level parameter. Mirjam provided many helpful comments and was particularly involved in the experimental setups. The content of this chapter has also been published as Dall et al. (2014a) and Tomalin et al. (2015) and is here presented in rewritten form.

9.2 Introduction

In the previous chapters, methods for producing filled pauses (FPs) more like those found in actual data were presented, and an improvement in perceptual quality was shown. But while we might be able to produce convincing FPs (more on how they compare in a psycholinguistic setting in Chapter 10), we do not know when to use them. That is, in normal TTS input text, FPs are not used, so while we can synthesise them well in applications in which the text can be expertly pre-prepared, e.g. well-crafted dialogue systems or set phrases in personal assistants, we cannot enhance text that does not contain them. This chapter will focus on just that – predicting where to insert FPs in to the text stream of an utterance.

To date, there have been only a few attempts at inserting FPs in speech synthesis systems. For predicting when to use FPs, Adell et al. (2007b) used a combination of n-grams and decision trees trained on a 317,000 word corpus to predict FPs. They obtain a high F-score, however the possible insertion points (IPs) were limited to those occurring after the 20 words most commonly followed by an FP – significantly simplifying the task. These kinds of distribution patterns should be well-modelled by n-gram language models (LMs) as FPs are very common, n-grams model common phenomena well and the method is limited to only the 20 most common word contexts. Andersson et al. (2010a) combined n-grams and the Viterbi algorithm in a lattice setting to find the best possible IPs of fillers and discourse markers using a limited training set of 2120 sentences: these were the transcriptions of the data used for training the actual synthesis system. The method of Andersson et al. (2010a) is geared toward picking examples which exist in the limited training data, something which is especially important in concatenative synthesis, but which limits the domain of the method. Ideally we would like an insertion system that is, to a large degree, independent of the voice used. One could argue that some speaker dependency is desirable as some speakers will use some FPs more often than others, but for the methods here discussed the focus is generally on predicting whether and where to insert an FP.

In this chapter, a method utilising a very large corpus of spontaneous speech (see Section 9.5) for FP prediction is presented. Focusing on the issue of predicting when to use an FP, and later also which FP to use, several experiments are presented. The first experiment determines whether such a corpus can act as a gold standard for FP prediction and whether people's predictions about FPs are consistent with reality and with each other. Secondly, the claim that there are 'right' and 'wrong' places to insert FPs is tested. By comparing synthetic sentences with either an FP inserted at the most frequently used IP from the first experiment, an FP inserted at unused positions from the first experiment or no FP at all in the sentence, one can find out if the position of an FP matters perceptually.

Once it has been established that there is indeed a pattern to predict, results of an initial system for FP insertion prediction is presented using n-gram LMs, a recurrent neural network (RRN) LM, support vector machines (SVMs) and decision trees (DT). Following that, an improved model is presented which is capable of inserting a variable number of FPs – dependent on a disfluency parameter – and a variety of different FP types – 7 different FP types – in a lattice framework.

9.3 Human Filled Pause Insertion

It is uncertain whether naturally occurring FPs can represent a gold standard for automatic methods to predict. There may not be any discernible pattern(s) to predict, and although the corpus studies reviewed in 4 suggest regularities in their appearance – these may not be strong enough patterns for reliable prediction. One way to elicit a measure of the predictability of FPs is to ask humans to insert FPs into sentences, and then compare their insertions to actual naturally occurring instances. By asking participants to insert FPs in sentences, several different questions can be answered:

Q1 Do humans agree on where to use FPs?

Q2 Do different data types elicit different agreement patterns?

Q3 Are predicted FP insertions in agreement with actual usage?

If the answer to (Q3) is positive, then it is possible to use data extracted from transcribed corpora as a gold standard. As multiple IPs may be valid in any given sentence, if the answer to (Q1) is yes, then this would allow for the gathering of human task performance data and metrics from informants which would provide multiple potential IPs for the automatic methods to predict and this human performance could set the bar for success. If (Q2) can be answered positively then it informs whether one should be careful about which domain FPs are applied to, whether sentence type affects FP use, and thus whether differing models should be applied in different scenarios.

9.3.1 Materials

Sentences were selected from two different corpora: the Wall Street Journal (WSJ) corpus (Paul and Baker, 1992) and the AMI corpus (Carletta, 2007). The WSJ data comprises news texts from the Wall Street Journal and the AMI data is spontaneous speech from multi-speaker meetings transcribed and divided by speaker. The two data types were chosen to allow for the investigation of whether subjects behave differently for spontaneous speech vs written news texts. Furthermore, for both AMI and WSJ sentences, a set of 15 isolated sentences and 15 paragraphs was selected, a sentence in the middle of each paragraph was then identified as the target sentence for people to insert FPs into. This allows for investigation into whether context affects people's choice of IP (Q2). The inclusion of the AMI data facilitated the use of sentences already containing FPs and made it possible to compare the predicted IPs to actual use (Q3). Each sentence contained at least one and maximally five FPs (mean = 2.9) and were required to be well-formed containing no other types of disfluencies. When presented to participants each sentence was stripped of all FPs when presented to participants. In total, 60 distinct sentences were used. Table 9.1 shows a few sample sentences.

9.3.2 Method

72 paid native English University of Edinburgh students were recruited. The 60 sentences were divided into two sets with equal amounts of text and sentence types. Each participant rated one of the two sets with sentences presented in a random order. The participants were instructed to imagine they were saying the sentence in a conversation, and then determine where they would be most likely to insert an FP. They were told to insert at least one FP, but were free to insert several if they thought it was more natural. The possible IPs were at any point between the words in the sentence, including the beginning or end.

9.3.3 Results and discussion

Due to experimenter error one AMI sentence contained an FP when presented to subjects and was excluded from the analysis. In order to determine if a predictable pattern exists and if it is non-random, a chance category was created. Using the overall statistics of potential IPs and mean number of inserted FPs, a simulation of the experiment was run 10,000 times in which FPs were inserted at random IPs in each sentence to find the chance values presented in Table 9.2, which also gives statistics of the data and shows subjects' agreement results.

With regards to (Q2) we can see that subjects insert FPs in a similar way regardless of the type of data they are faced with: paragraph or isolated sentences, news or spontaneous text. The AMI and WSJ differences are the largest, but still insignificant. This difference is possibly due to the lower number of potential IPs in the WSJ sentences which naturally result in a slight agreement increase. Focusing then on the overall results; subjects are quite consistent compared to each other, as the most frequently used IP represents 26.35% of all insertions and the top three IPs represent 56.49%, which is substantially above chance levels at 9.54% and 28.61%. We also see that 22.07% of IPs are never used compared to 2.77% by chance. This demonstrates that (Q1) seems to be the case, people generally agree on where it is acceptable to use an FP and they also agree that there a places where *not* to use an FP.

	Example Sentence
WSJ Isolated	“It is important for actors to reinvent themselves.” “The proposed pact was first reported in the Washington Times.”
WSJ Paragraph	“The mayor said he was driving when he saw smoke in the direction of the building. **He turned on his police scanner and heard there was an explosion.** Heydt said he arrived before the fire trucks.” “Still, it is astonishing to count the number of people at the White House who are tickled to death with the book. **Lord knows they all had their say, got to babble to a guy who was a character in a real movie.** On balance, the White House seems positively relieved.”
AMI Isolated	“I think it’s a multiple chip design and it’s maybe printed on to the circuit board.” “It was discussed in the last meeting which was opened by the presentation from the interface.”
AMI Paragraph	“So it has to be simple another point is we have to skip the teletext **because in the world of upcoming internet we think teletext is going to be a thing of the past** and it’s a function we don’t need in our remote control.” “And we asked them how relevant they think the buttons are **the power volume and channel selections are very relevant** teletext is less relevant but also important.”

Table 9.1: Examples sentences from the WSJ and AMI sentences, AMI sentences were presented without the original FPs. For the paragraphs the bold-faced part was the target sentence to insert FPs into.

Condition	Possible IPs	Used IPs	Ins FPs	Top Used IP	Top 3 Used IPs
WSJ	12.77	80.68%	1.40	28.07%	59.14%
AMI	16.48	75.73%	1.67	24.86%	54.19%
Isolated	14.59	77.54%	1.51	26.78%	58.17%
Paragraph	14.60	78.31%	1.55	25.98%	54.91%
All	14.59	77.93%	1.53	26.35%	56.49%
Chance	14.59	97.23%	1.53	9.54%	28.61%

Table 9.2: Mean values over all sentences for Possible IPs, Used IPs, Inserted FPs and agreement of most used and three most used IPs. IP = Possible insertion point in sentence. FP = Filled pause. Ins = inserted. The percentage values for Chance is the mean value obtained for a simulated trial over 10000 trials.

Comparing the actual IPs from the AMI data to the manually chosen IPs, we find that for 40.3% of the sentences there was a match between an IP in the original data and the *most frequently chosen* IP in the test data. This is compared to a lower internal consistency in the manually chosen IPs of 24.86%. Almost all (96.6%) of the original AMI sentences had an FP in one of the three most likely chosen IPs, compared to a 54.19% internal consistency, and only four (4.82%) of the original IPs were not predicted at any time in the AMI test data. This demonstrates a very good consistency between subjects' predicted usage and their actual usage (Q3), and a good consistency in subjects' predictions (Q1). We can therefore use these values as a guide to compare automatic methods against when using transcribed spontaneous speech as a gold standard.

9.4 The Perceptual Effect of FP Position

While humans are reasonably consistent in where they use, and predict, FPs, this does not demonstrate that the IP makes a perceptual difference to listeners. In speech synthesis it is generally accepted that incorrect pausing is detrimental to the processing of speech synthesis (e.g., Taylor (2009), p.142), however, there is, to my knowledge, only one paper that has investigated this. Scharpff and van Heuven (1988) measured the effect pausing has on intelligibility of low quality speech synthesis. They conclude that the intelligibility of low quality speech improves when pauses are inserted at prosodic boundaries, but deteriorates when other locations are chosen. It is likely that an FP may behave similarly, that there are ‘right’ and ‘wrong’ places for FPs, however this has not been tested (e.g., Andersson et al. (2010a) assumed this to be true).

The objective of this second experiment is therefore to analyse whether there are IPs where one should not insert an FP and if well-placed FPs are preferred over no FP insertion.[1]

9.4.1 Materials

Twenty of the thirty AMI sentences from the human insertion experiment were used in the perception experiment. FPs (in this case ‘UH’) were inserted either at the most likely place (according to the judgements from the previous experiment) or randomly in one of the unused IPs (i.e., an IP that wasn’t chosen by any of the participants in the insertion experiment). The sentences were synthesised using a female voice based on HTS 2 (Zen et al., 2007a) and about 8 hours of read speech, in a system that was newer than, but broadly similar to, that in Yamagishi and Watts (2010), which is representative of state-of-the-art HMM-synthesis (this is the same system as used in the psycholinguistic experiments of Chapter 4). During synthesis, the ‘UH’s were treated as regular word tokens in the input stream, as argued for in Clark and Fox (2002) and Andersson (2013). However, as discussed in the previous chapter, producing spontaneous speech elements using a voice based on read speech results in less natural sounding synthesis than if it was based on spontaneous speech. To circumvent the worst quality problems with the synthetic ‘UH’s, a selection was made in which a good sample of a synthetic ‘UH’ was identified in a number of synthesised utterances and this ‘UH’ was spliced into the synthesised sentences at the appropriate locations. Experimental materials available at the thesis repository (Dall, 2017).

9.4.2 Method

The listening test was conducted through Amazon Mechanical Turk (AMT). Two conditions were created, each consisting of 20 sentence pairs. Pairs were presented in a random order, and the comparisons within a condition were either FP in top position versus FP in an unused position, or FP in top position versus no FP. The task for the Turkers was to listen to the two sentences and choose which version they preferred, if

[1] Please note that in chronological terms this experiment was performed prior to the improved FP synthesis presented in the previous chapter, and thus uses a standard read speech-based voice. See Chapter 10 for further evaluation of the improved FP synthesis.

Top	Unused	No FP	No Preference
7.5%	-	**85.3%**	7.3%
61.0%	29.3%	-	9.75%

Table 9.3: Preference scores between synthetic sentences with either an FP in the most used (Top) IP, an FP in an unused (Unused) IP or no FP in the sentence (No FP). Top and Unused IPs were determined in Section 9.3

any. The instructions were: "You will be listening to pairs of sentences. Please choose which one you think sounds most natural". The options were: "Sample 1", "Sample 2" or "No preference". It was requested that only native speakers of English carried out the work. Quality control is an issue with AMT; to overcome this, the work was only offered to Turkers with master worker status. Master status indicates a worker has previously completed work to a high level of satisfaction. This criterion should ensure that only workers who would carry out the task diligently did the task, as they would not want to risk their master worker status. In addition, we included three safety questions in which the Turker was instructed, in the audio of the safety question, to select a certain option. Turkers that failed to respond correctly were excluded under the assumption that they had not paid enough attention to the test. 44 Turkers completed the work, four were excluded as they failed the safety questions. In total, responses from 20 workers per condition were considered.

9.4.3 Results and discussion

All "No Preference" results were evenly divided between systems to allow for statistical significance testing using the exact binomial test. Table 9.3 shows the results of the listening test. Listeners have a clear preference for FPs inserted in the top IP ($p < 0.001$) compared to FPs inserted at a random unused position. However, when given the choice between a sentence containing a top FP versus a sentence without an FP, listeners overwhelmingly choose the fluent sentence as the more natural ($p < 0.0001$). This is not surprising since FPs are often considered disfluencies, and these are generally judged to be undesirable if naturalness is equated with formal correctness (Christenfeld, 1995). However, as mentioned throughout the thesis, FPs are, in practice, very "natural" since they are prevalent in spontaneous speech and serve a range of interesting purposes.

The results seem to be the opposite of those found by Adell et al. (2007b) where listeners found sentences with FPs more natural than sentences without. A major difference between the current experiment and Adell et al. (2007b) is the way the question was framed. Listeners in Adell et al. (2007b) heard pairs of sentences with and without FPs and were asked whether the FP increased the naturalness of a voice for a dialogue system. Their focus was drawn to the FPs explicitly and the question was further framed by specifying the style of speech. In the current study, it was purposefully not specified that the sentences contained FPs and were from conversational speech, as it was felt that might prime the participants towards choosing FPs – particularly in the light of Christenfeld (1995) and the results in Chapter 3 which suggest that explicit style focus and instructions strongly influence the resulting judgements.

Another possible reason for the difference is that Adell et al. (2007b) used a concatenative system in which they hand-picked samples of actual FP recordings based on their earlier work in Adell et al. (2006). By contrast, our system used a voice trained on read speech containing no FPs. The FPs in Adell et al. (2007b) most likely sound more natural than those used here, despite the selection and splicing, and our results may partly reflect poor synthetic FP quality. This leads to the conclusion that while there are "right" and "wrong" places to place an FP, and these places conform to human usage, it is not enough to simply insert FPs in the right places, they also have to sound right if they are to be acceptable to a listener.

9.5 Automatic Filled Pause Prediction

The human insertion experiment indicates regularities we can predict, and the above experiment that both quality and position of the FP is important. Focusing on the IP, various techniques for automatic FP insertion are explored here. In order to facilitate the use of predictive techniques, a training data set was defined using data from the AMI (Carletta, 2007), Fisher (Cieri et al., 2004) and Switchboard (Goodfrey et al., 1992) corpora and also an unreleased corpus of British conversational telephone speech. All these contain transcriptions of spontaneously produced speech, thus containing FPs, and can be used for modelling. As each of the different corpora uses slightly different transcription conventions, they were all preprocessed to normalise markup of, particularly, "disfluent" elements. In general, markup was removed and all disfluent elements were treated as a word token in the text stream, except for nonverbal sounds. That is, FPs, repetitions, restarts, foreign word use and other verbal "disfluent" elements were stripped of any markup and kept as word tokens. However, laughter, breathing, grunts and the like, if marked, were removed. In total this set contained 1,164,938 sentences and a total of 19,467,756 word tokens – a far larger corpus than previously used. The two most common kinds of FPs ('UH' and 'UM') were mapped to a single type, 'UH', since in this initial experiment the focus was primarily on finding the most likely IP irrespective of FP subtype (see Section 9.6.1 for predicting individual FP types). Sentences containing fewer than two words were removed because backchannels were not of interest and were frequent enough to potentially affect prediction. Development (dev) and test sets were defined using the same corpus. They each contained 2000 sentences, 1000 of which contained FPs and 1000 without, and consisted of 35,131 and 35,100 words respectively. The FP-containing sentences were chosen to be similar to the sentences used in the human insertion experiment: word length was restricted in a similar way, and they contained exactly three FPs. The choice of three FPs was due to this being close to the average number of FPs in the real sentences used earlier (which was 2.9), and because it allows comparisons to the top three used IPs from the human experiment, as such, systems inserting a single FP can be scored and compared to that. Using the training data, six automatic FP insertion systems were built:

- *Random*: Randomly inserts a single 'UH' into a sentence.
- *N-gram LM*: A standard 4gram language model (LM) was built using the SRILM toolkit (Stolcke et al., 2011) from the training data (68K wordlist, Kneser-Ney discounting).
- *Recurrent Neural Network LM*: A Class-based Recurrent Neural Network (RNN) LM was built using the RNNLM Toolkit (Mikolov et al., 2011) from the training data. The RNN was 500 neurons wide and, for speed reasons, was trained using 250 classes.
- *Interpolated RNN and N-gram LM*: The N-gram and RNN LM scores were linearly interpolated on a by-word basis to re-rank the potential sentences.
- *Support Vector Machine-based (SVM)*: A vector of features was extracted for each IP in each sentence in the training data:
  - syllable count of word following IP
  - phrase boundary associated with IP
  - clause boundary associated with IP

  - 4gram log probability for sentence with UH in IP
  - Part-of-Speech of word following IP

- *Decision Tree-based*: A CART-style Decision Tree (DT) was built using R[2] and the same features as for the SVM above. The tree was pruned by selecting the complexity parameter associated with the smallest cross-validated error.

For the extracted features, the syllable count was obtained using tsylb;[3] phrase boundary, clause boundary and part-of-speech were obtained using the Stanford Parser,[4] and the 4gram scores from the N-gram LM. All features were scaled and normalised so they could be expressed as floating point values between log(0) and log(1). SVM models were built for all possible feature combinations using SVM-Perf.[5]

Note that the n-gram and RNN systems and the DT and SVM systems represent two differing approaches to the problem, using two differing types of models. The first treat the problem as a language modelling problem by modelling the probability of seeing a string of words given a training corpus. As language models, these models only rely on the word tokens in the training corpus as their input. The DT and SVM, on the other hand, are classifiers, and so inherently treat the problem as one of classification. These models rely on a feature set (described above) to guide the classification which is used to assign a probability to each potential sentence. There is thus an inherent incomparability in the feature set used for the models, just the word strings vs. a designed set of features, but as the models are of differing types this is acceptable. It is not the intention here to test which is the superior type of machine learning model, merely to try and establish which models perform the best on this task. Note that if the RNN was utilised as a classifier instead of a language model, it would be problematic if it was not given the same set of features as the SVM and DT, this is not the case so we can talk about the models effectiveness for this task.

For each system insertion was done by creating a list of each possible sentence with an FP in one IP, and also the sentence with no FP at all. These sentences were then scored and the highest probability sentence chosen. Outputs were produced for all systems, and they were scored using precision, recall and F-score. All systems predict either one FP or no FP. As the FP containing sentences in the test and dev sets contains three FPs, a 'correct' prediction therefore occurs when the system predicts no

[2] http://www.r-project.org

[3] ftp://jaguar.ncsl.nist.gov/pub/tsylb2-1.1.tar.Z

[4] http://nlp.stanford.edu/software/lex-parser.shtml

[5] http://www.cs.cornell.edu/people/tj/svm_light/svm_perf.html

FP when there is none or correctly predicts an FP in one of the three IPs containing an FP in the dev/test sentence. Precision is defined as correct if an FP is predicted and it is in a correct position. Recall is when an FP should be predicted and one is predicted. F-score is the amalgamation of the two scores. As mentioned above, this situation is similar to the 3-best results of the human insertion and thus a (cautious) comparison can be made.

9.5.1 Results and discussion

From Table 9.4 we can see that the best performing system is the RNN/N-gram interpolation. It is clear that the RNN and N-gram LMs complement each other. Although the N-gram is conservative in its prediction of FPs (only 359 of 1000), it is much more precise (42% precision) than the other systems. By comparison, the RNN massively over predicts the number of FPs (1877 instead of 1000) and therefore is not as precise (32%), but it has a much better recall (52% vs 16% for the n-gram). By interpolating the two, we predict a number of FPs much closer to the actual 1000, namely 1217, and obtain the best precision (50%) and second best recall (47%), yielding the highest F-score (0.48). It is likely that the reason for this difference lies in the way the two LMs work, where the 4-gram will only output an FP when there is sufficient local evidence in the 4 word history, the RNN is capable of considering longer-range dependencies. This is important since, e.g., sentence length has an impact on the likelihood of an FP being used (longer sentences contain more FPs).

Both the SVM and DT perform disappointingly, with the simple 4gram LM performing as well as the best SVM and also being the SVM's most useful feature. Whereas the DT achieves the best results with all features, features (a)-(c) and (e) only confuse the SVM, presumably because they are not context-dependent. This is in stark contrast to the DT performance in Adell et al. (2007b) which was much better than the results found here. The difference is likely due to Adell et al. (2007b) limiting the number of possible IPs to those 20 words most often followed by an FP, which simplifies the task significantly. Furthermore, it is possible that the utilised feature set was sub-optimal, and the inclusion of other features may have improved the performance of the models.

The performance of the RNN/4gram system is encouraging as it shows we can quite reliably predict where to insert FPs in text, and the performance is on par with that of the human top-3 performance (56% vs 54%). If we allow the system to produce

System	Precision	Recall	All F	UH F
Random				
dev	13%	16%	0.14	0.16
test	14%	17%	0.15	0.18
4-gram				
dev	49%	15%	0.23	0.26
test	48%	16%	0.24	0.27
RNN				
dev	31%	**51%**	0.39	0.51
test	32%	**52%**	0.40	0.53
RNN/4-gram				
dev	**53%**	**51%**	**0.52**	**0.57**
test	**50%**	47%	**0.48**	**0.54**
SVM All				
dev	24%	17%	0.20	0.20
test	26%	16%	0.20	0.19
SVM Best				
dev	27%	22%	0.24	0.27
test	29%	23%	0.25	0.27
DT All (Best)				
dev	25%	16%	0.19	0.19
test	25%	17%	0.20	0.21

Table 9.4: Overview of the 1-best output, the best system scores are bold-face. ‘All F’ is the F-score when considering the full dev/test set. ‘UH F’ is when only considering sentences containing FPs. ‘All’ refers to the system using all features. ‘Best’ refers to the best performing feature combination.

a 3-best list, the precision (85%) and recall (81%) improves even further (F=0.83), demonstrating that reasonable IPs are being identified.

In this section, the focus has been on the task of identifying IPs for FPs in different kinds of data (e.g., conversational speech and written news text). The results of the human insertion experiment demonstrate that the type of data does not affect results, that there is good consistency between human subjects' predictions of FP usage and their actual usage, and also a good agreement between predictions from different people. The perceptual test of high-use vs. no-use IPs shows that the IP of an FP also has an effect on peoples preferences.

Overall, this confirms that FP insertion is not merely random, and therefore it can be modelled in speech synthesis systems. As an initial step towards this, the performance of various automatic systems has been compared and contrasted, and it was shown that an interpolated n-gram and RNN LM produced the best output. The superior performance of this system suggests that the accurate modelling of local and long-range lexical and syntactic contexts is central to this task, and the systems described here could be improved by the inclusion of additional features or complementary modelling methods (e.g., dependency grammars). However, considering the well performing n-gram/RNN LM system, this is outside the scope of this thesis. It is also possible to extend the methods summarised here to the task of inserting other kinds of spontaneous speech phenomena such as repetitions, restarts and discourse markers and also to the use of potentially more than one FP in each sentence.

9.6 A Lattice-based Approach to Automatic Filled Pause Insertion

As the previous experiments demonstrate, the insertion of an FP into the text stream can be well predicted using an interpolated n-gram/RNN LM. However, the presented method pooled the use of the two most common FPs – 'UH' and 'UM' – and predicted maximally one FP. In this section, that approach will be extended to the insertion of multiple FPs, and multiple types of FPs, in multiple IPs. Therefore the sentence "I never liked games" could be modified automatically to become "**UM** I never liked **UH** games". In addition, a disfluency parameter (DP) that determines the degree of disfluency in the output text is also introduced. The DP takes a value in the range $[0,1]$, where 0 = maximally fluent and 1 = maximally disfluent. Finally, while the previous method used linear interpolation of word-level N-gram and RNN LM probabilities to re-rank the potential sentences, a more robust lattice-based rescoring method is introduced here. As a modelling technique, it has clear advantages since simple re-ranking strategies become computationally inefficient when multiple FPs can be inserted in multiple IPs.

9.6.1 Lattice-based LM Prediction with a Disfluency Parameter

The lattice-based FP-insertion system presented here is similar to those recently implemented for Automatic Speech Recognition (ASR) tasks in Chen and Manning (2014); Liu et al. (2014). In the context of language modelling for ASR, RNN LMs have become increasingly popular in recent years due to their inherently strong generalization performance – as also evidenced by the effectiveness of the method in the previous section. Specifically, Chen et al. (2014) has shown that full output layer RNN (f-RNN) LMs facilitate an efficient parallisation of training in a Graphics Processing Unit (GPU) implementation (as opposed to the CPU and class-based RNN approach used in the previous section). In addition, when used in a lattice-rescoring framework, they give both perplexity and Word Error Rate (WER) improvements over standard RNN LMs. This is due in part to their use of an unclustered "full-output" architecture.

This framework can be adapted for the FP-insertion task. There are five main stages in the modified process:

1. Create initial lattices in which each FP is accessible from each word token (Figure 9.1).
2. Expand the initial lattices using an n-gram (6g).
3. Rescore the expanded lattices using an interpolated LM with a linear weighting of the n-gram and f-RNN LMs.
4. Output an n-best list for each sentence (where $n = 10000$).
5. Specify the desired degree of disfluency using the DP and generate final 1-best disfluent output.

The use of an n-best list is done in order to use the DP to control the desired level of output disfluency as this cannot be done directly through the lattices. After lattice decoding, the versions of sentence S in the n-best list will have varying token counts

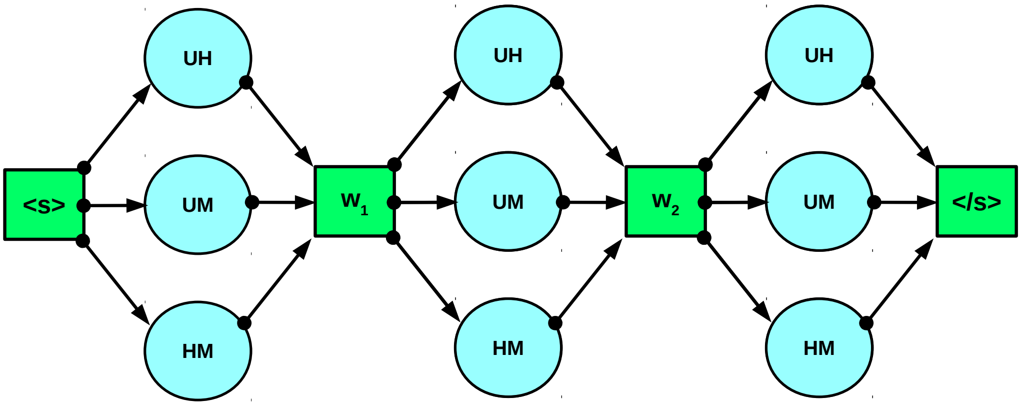


Figure 9.1: Example Initial Lattice for two words and three FP types.

DP	Output Sentence
0.00	WELL I GUESS THEY WERE SAYING
0.25	WELL I GUESS THEY WERE SAYING **UM**
0.50	**UM** WELL **UH** I GUESS THEY WERE SAYING **UM**
0.75	**UM** WELL **UH** I GUESS **HM** THEY WERE SAYING **UM**
1.00	**UM** WELL **UH** I GUESS **HM** THEY **UH** WERE SAYING **UM**

Table 9.5: An example of the impact of DP values on output token sequence

since they will contain different numbers of automatically inserted FPs. All versions of S with p tokens are rank-ordered using the sentence-level interpolated LM score, and the 1-best version is output. The DP then determines which of these 1-best output is picked – if the DP is 0.0 the first sentence is taken (the one with least FPs), if it is 0.5 the middle one and if it is 1.0 the last one – and so on. This provides the DP that determines the degree of disfluency. Note that this does not mean that using 0.0 will always yield no FPs and 1.0 always yield an FP in all IPs, if no sentence with any given number of FPs occurs in the initial n-best list (of 10000 in size) then a sentence containing that many FPs will not be available for selection. The impact of varying the DP parameter is shown in Table 9.5. This example provides a concrete instance of the impact that the DP value has on the resulting token sequences, and it illustrates the graded nature of the different outputs.

	Train	**Dev**	**Test**
UH	213,924 [1.09%]	3660 [2.29%]	3658 [2.44%]
UM	200,499 [1.02%]	3331 [2.09%]	3392 [2.26%]
OH	123,028 [0.63%]	2035 [1.27%]	2083 [1.39%]
AH	69,288 [0.35%]	1053 [0.66%]	348 [0.23%]
UHUM	29,515 [0.15%]	432 [0.27%]	423 [0.28%]
UHU	16,180 [0.08%]	222 [0.14%]	228 [0.15%]
HM	3,456 [0.01%]	61 [0.04%]	55 [0.04%]

Table 9.6: FP occurrence counts and % of data for the Train, Dev and Test sets

9.6.2 Data

The LMs used in the experiments were trained on the same corpus as in the previous section, which contains roughly 20M words/1M sentences. Dev and test sets were extracted from the same corpus, and they comprised 7,365 sentences (145k words) and 6,910 sentences (139K words) respectively, with no overlap between them. The dev and test set were different, and larger, than those in the previous section. This was done in order to facilitate a distribution of each FP modelled similar to that of its occurrences in the data. Each sentence in the dev and test sets contained at least one FP, and these FPs were removed to create the 'fluent' version of the test sets that were processed by the FP-insertion systems. Seven different FPs were modelled by the various FP-insertion systems: 'UH', 'UM', 'OH', 'UHUM', 'UHU', 'HM', and 'AH'. Information about the occurrence of these FPs in the training data is given in Table 9.6. As not all sentences in the training data contain FPs, the overall amount of each type of FP in the dev/test sets is larger than in the training data – each occurs about twice as frequently as a consequence of about half the training sentences containing at least one FP – but, crucially, the distribution is similar.

As the counts in Table 9.6 indicate, the FPs 'UH' and 'UM' occur most frequently in the training data. The fact that some of the other FPs have relatively low counts (<30,000) facilitates the exploration of the impact of data sparsity on the modelling of speech disfluencies.

9.6.3 Insertion Experiments and Results

Three FP-insertion systems were compared:

1. **N-gram**: a standard 6gram LM built using the training data; SRILM toolkit Stolcke et al. (2011); Kneser-Ney discounting

2. **f-RNN**: a non-class-based f-RNN LM with 512 hidden layer nodes

3. **N-gram+f-RNN**: the 6gram and f-RNN LM interpolated at 50%-50% weighting

For the n-gram system, the choice of using a 6gram instead of a 4gram was made based on the observation that the previous 4gram seemed to make very local decisions, and it was the hope that by increasing the history length an improved recall could be obtained with the n-gram. The initial lattices (Figure 9.1) were expanded and rescored using the n-gram, the f-RNN LM, and the interpolated n-gram+f-RNN LMs. System performance was evaluated using the precision, recall, and F-score metrics. Precision here means if an FP was predicted it was the correct FP in the correct IP – recall if an FP of a particular type should have been predicted and one was. F-score is the standard combination of precision and recall. For the interpolation of the n-gram and f-RNN LMs, the total probability of a sentence (*sentenceProb*) is a linear interpolation of the LM probability scores for each sentence for the n-gram (*ngramProb*) and f-RNN (*rnnProb*) using a separate weight (*ngramWeight* and *rnnWeight*) for each LM:

$$sentenceProb = ngramWeight * ngramProb + rnnWeight * rnnProb \tag{9.1}$$

such that $ngramWeight + rnnWeight = 1$.

The range of interpolation weightings was explored for the n-gram+f-RNN LM system at each 10% interval, and the 50%-50% weighting gave the optimal performance on the dev set. Consequently, the 50%-50% weighting was adopted for all the experiments reported. The metric scores were also used to determine the optimal DP value based on the dev data and Figure 9.2 shows the scores for the n-gram+f-RNN LM system when varying the DP value. An inverse relationship between precision and recall is apparent, and a DP value of 0.5 achieves a desirable balance between these extremes – F-score also stabilises at this point. Similar patterns were obtained for all three systems, so the DP was set to 0.5 for all subsequent experiments.

Table 9.7 shows that the n-gram+f-RNN LM system obtained the best (sometimes joint best) precision and recall performance for every case except the recall results for the Test set. Notably, the n-gram+f-RNN LM system obtained the best F-score results for both the dev and test sets. This suggests that the interpolated system combines the

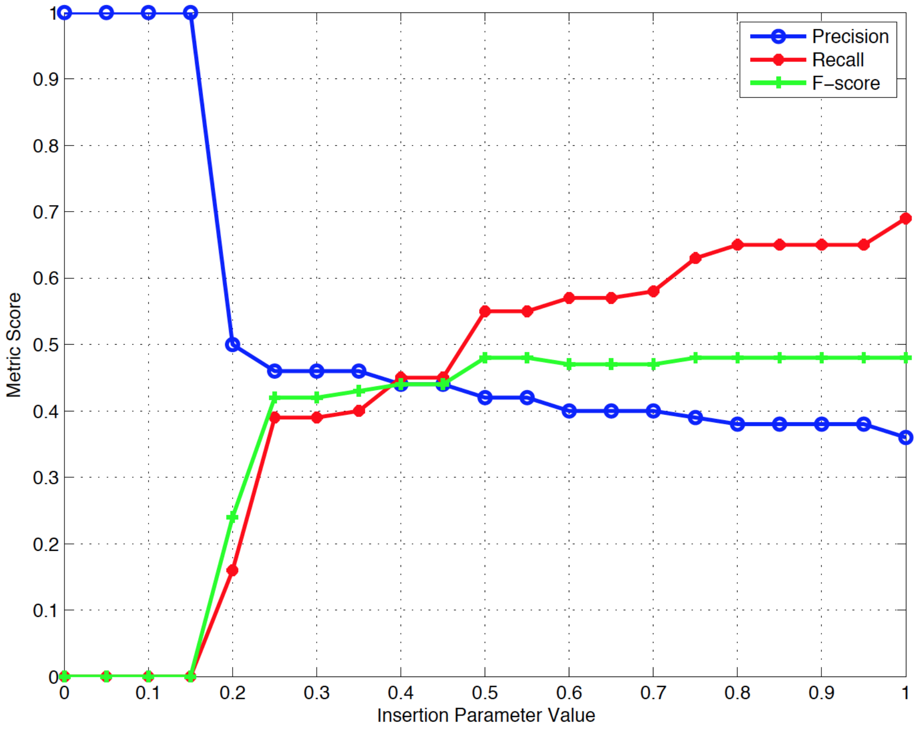


Figure 9.2: Precison, recall, and f-score for n-gram+f-RNN LM for Different DP Values

	Precision		**Recall**		**F-score**	
	Dev	**Test**	**Dev**	**Test**	**Dev**	**Test**
N-gram	0.41	0.44	**0.55**	**0.60**	0.47	0.51
f-RNNLM	0.42	**0.47**	0.52	0.57	0.47	0.51
N-gram+f-RNNLM	**0.43**	**0.47**	**0.55**	0.59	**0.48**	**0.52**

Table 9.7: Dev and Test results for the N-gram, f-RNN, and N-gram+f-RNN LM systems using a 50/50 interpolation level and a DP value of 0.5.

	Dev		**Test**	
	Ref	**Hyp**	**Ref**	**Hyp**
UH	3660 [2.29%]	6359 [3.97%]	3658 [2.44%]	6311 [4.10%]
UM	3331 [2.09%]	4711 [2.94%]	3392 [2.26%]	4201 [2.73%]
OH	2035 [1.27%]	3685 [2.30%]	2083 [1.39%]	3538 [2.30%]
AH	1053 [0.66%]	192 [0.12%]	348 [0.23%]	209 [0.14%]
UHUM	432 [0.27%]	73 [0.46%]	423 [0.28%]	92 [0.06%]
UHU	222 [0.14%]	8 [0.00%]	228 [0.15%]	23 [0.01%]
HM	61 [0.04%]	2 [0.00%]	55 [0.04%]	0 [0.00%]

Table 9.8: Number of occurences of each FP in the dev/test sets (ref) and the N-gram+f-RNNLM system output (hyp).

complementary properties of the two component LMs. The is, the n-gram+f-RNN LM system is comparatively more robust than either the n-gram or f-RNN LM systems, and the latter two are beneficially interpolated in the lattice-based framework.

It is worth noting that the n-gram system in this case achieved the highest recall scores – and the RNN LM a higher precision – which is the opposite of earlier. This is may be partly due to the longer history (6 vs 4) of the n-gram. However, the main reason is the DP parameter which encourages a certain amount of FP insertion – stabilising the number of predicted FPs and thus helps recall naturally for the n-gram by making it predict more FPs and improves precision of the RNN by making it predict less FPs (but with a higher precision).

Table 9.8 gives the number of occurrences for both the dev and test sets and the n-gram+f-RNN LM system output hypotheses. These counts show that the n-gram+f-RNN LM models the various FP subtypes rather differently. There is a tendency to overgenerate the three most frequently occurring FPs (i.e., 'UH', 'UM', 'OH'). By contrast, the system undergenerates the less frequently occurring subtypes (e.g., 'UHU', 'HM'). Presumably this is a consequence of the occurrences in the training data, which ensure that the LMs associate higher likelihoods with frequently occurring FPs – and these higher likelihoods may well have the consequence of dominating rare FP subtypes. The prediction patterns for all FP subtypes are similar for the dev and test sets. Table 9.9 further demonstrates this by giving the precision, recall, and F-score scores for the distinct FP subtypes.

The scores show a fair amount of variation between the different FP subtypes. The

	Precision		**Recall**		**F-score**	
	Dev	**Test**	**Dev**	**Test**	**Dev**	**Test**
UH	0.42	0.46	**0.72**	**0.74**	0.53	0.57
UM	0.42	0.45	0.56	0.54	0.48	0.49
OH	0.48	0.53	0.70	0.70	**0.57**	**0.60**
AH	0.14	0.12	0.03	0.08	0.05	0.10
UHUM	0.35	0.58	0.04	0.09	0.08	0.16
UHU	0.25	**0.81**	0.01	0.06	0.02	0.11
HM	**0.50**	0.00	0.02	0.00	0.04	0.00

Table 9.9: Individual FP Results for the N-gram+f-RNN LM system. Bold faced are the highest scores in each category.

scores for the three most frequently occurring FPs are relatively stable across the dev and test sets, achieving F-scores in the range 0.48-0.60. By contrast, the scores for the less common FPs sometimes fluctuate considerably. This quantifies the impact of the data sparsity shown in Table 9.6 – low probability FPs are badly modelled, however, high probability FPs are much better modelled.

All in all, the system presented here achieves a high degree of accuracy when predicting FPs. The overall F-score of 0.52 for the test set is similar to that achieved in the 1-FP output systems previously presented, despite the task being harder. This demonstrates that multiple FP types can be usefully predicted through a lattice-based system controlled by a disfluency parameter. This, at the same time, allows for graded FP output – with the potential for simulating, e.g., speaker certainty about a specific topic.

9.7 Overall Conclusions

In this chapter, an approach to FP insertion modelling has been presented. An initial investigation into the consistency of human filled pause insertion revealed that human predictions are consistent with actual usage, and that FP position matters perceptually. Initial methods for the insertion of up to one FP in a sentence were presented. In a comparison between a 4gram LM, RNN LM, an interpolated 4gram/RNN LM, a decision tree and a SVM-based approach, the interpolated 4gram/RNN LM performed best, and very close to human performance (although this is a cautious comparison).

Extending the initial n-gram/RNN LM approach to multiple FP types and multiple FP insertions, a lattice-based rescoring framework has been presented. This framework enables multiple FPs to be inserted into multiple IPs. The experiments involving seven different FP types show that, using the F-score metric, the 6gram+f-RNNLM system is more robust than its constituent 6gram and f-RNNLM sub-components since it combines their complementary tendencies.

This system utilises a disfluency parameter which allows for gradeable FP output, allowing for potentially modifying the perceived mental state of the speaker. This will, however, not be explored in this thesis as it is considered outside its scope.

Chapter 10

Tool Release and Final Evaluations

10.1 Introduction

In this thesis, a number of different systems all attempting to tackle the use and synthesis of spontaneous data and phenomena has been presented. However, the methods for pronunciation variant forced alignment and synthesis have not been combined or utilised in the sections looking at specific FP modelling and linguistic feature set enhancement. This chapter will present a final evaluation combining all of these methods into one voice and comparing it to a voice without these improvements. Furthermore, a re-run of the reaction time experiment from Chapter 4 will be presented in which it will be tested if a voice utilising all the here presented improvements is capable of replicating the effects found in natural and vocoded speech.

Finally, a front-end, SiRe, developed for the purposes of this thesis will be described and is released under a liberal license together with the corpus of read and spontaneous speech used in this thesis.

10.2 Reaction Time Revisited

In the reaction time experiment presented in Chapter 4 it was found that vocoded, but not synthetic, speech could replicate the effects found in natural speech – namely faster reaction times to a target word in a sentence if the target word was preceded by a FP as compared to a SP (silent pause) of equal length. One potential reason given for this was bad FP realisation, another flat prosody. In this thesis a number of improvements to FP synthesis have been presented, along with a new linguistic feature set which improves naturalness and affects primarily F0 prediction – suggesting that phrase level prosody

	Baseline	**Proposed**
FP Phone	/V/	/UHV/
Data Marking	Read	Switch
Alignment	Standard	Variant
Pronunciation Choice	Standard	Fully Reduced
Linguistic Features	Standard	Position and Parsing

Table 10.1: The proposed and baseline voices used in this chapter. FP Phone refers to the phone representation of the filled pause of Chaper 7, Data Marking to the data marking technique of the same Chapter. Alignment to the forced alignment type described in Chapter 5. Pronunciation choice to the methods described in Chapter 6 and the linguistic features to the sets presented in Chapter 8.

is where the main improvement lies.

As these improvements coincide with the suggested reasons for the inability of synthetic speech to replicate the effects in natural speech, this will here be tested by re-running the RT (reaction time) experiment. As the initial experiment used a generic read speech-based voice not based on the corpus presented in Chapter 3 the RT experiment will be run using two different voices based on the read and spontaneous corpora from this thesis. One data marked voice using the standard FP phone from Combilex, the standard linguistic feature set, marking all speech as being read and using standard forced alignment and pronunciation choice – this is the baseline voice. The other voice is a data marked voice using the switching technique proposed in this thesis, the linguistic featureset including both PCFG and dependency parsing, utilising the specialised /UHV/ phone and pronunciation variant forced alignment with full reduction for pronunciation choice – this is the proposed voice of this thesis. See Table 10.1 for a summary of these.

The hypothesis that is tested is that the proposed voice will be able to replicate the effect found in natural speech. It is quite possible however, that the voice will behave no different than a standard voice, it may be these improvements are not enough. A secondary hypothesis is that even though the voice may not replicate the RT effect, it may still facilitate faster general reaction times over a standard voice due to the improvements proposed in this thesis.

10.2.1 Data and Method

The two voices were the ones described above and summarized in Table 10.1. The experimental sentences used were the same as in Chapter 4. That is, 79 critical and 37 filler sentences. In each of the critical sentences the target word was the first content word following the FP or SP, and in the filler sentences the target word was not immediately preceded by an FP or SP. Both target and filler sentences contained other FPs and SPs. All 79 critical sentences were synthesised with each voice with either an SP or FP preceding the target word as described and also all 37 filler stimuli with each voice. One filler and one critical sentence was reserved for a trial run. In total that gave 312 critical sentences and 72 filler sentences. The critical sentences were evenly divided into four groups of 78 sentences and the fillers into two sets of 36. Four sets of 78 critical and 36 filler sentences were then made, each with 38 SP and FP critical stimuli and 36 filler stimuli. One set of six trial sentences was also created (one from each condition). As one of the major differences in realisation between the two voices is the duration of the FP a decision was made to let each system realise FPs and SPs as naturally as they could, i.e., without modification. Although this means that the durations of the FPs and SPs across and within voices would vary, this was deemed the best solution. Alternatively one could have matched all durations, but this would remove one of the differences between the voices in FP realisation, and which duration model to pick as the matched is unclear. Matching FP duration across voices was problematic for the same reason, and matching FP and SP durations within each voice was also not desirable as we could then not compare overall RTs between voices as per the secondary hypothesis. Letting each system produce each phenomenon without attempting to match durations could arguably be a confounding factor, however, as the experimental setup has also been shown to work even when no pause is present (as in, e.g., Fox Tree (1995)) it is unlikely to be an issue. In fact by not manipulating any of the synthesised stimuli we test each system in as optimal conditions as possible. Another difference to the previous experiment is that the focus is solely on the FP 'UH', therfore, all repetitions and 'OH's included in the previous experiment were replaced with 'UH's or SPs as appropriate.

36 paid native English speakers were recruited to take part. The procedure and location was the same as in Chapter 4. That is, participants were seated in a soundproof booth in front of a screen wearing high quality headphones. They were first given a trial run after which they had the opportunity to ask clarifying questions. Next,

the subjects were given three equally sized parts to listen to, between each they had the chance to take a quick break. Each stimuli was presented by first a +-sign for 500ms followed by the target word for 1000ms and after 500ms the sentence was played. The critical measurement is the RT in ms of the participant from the word onset of the target word. Each participant was presented with one of the four sets of stimuli for a total of nine complete ratings of each set of sentences for each condition. This yielded 711 measures per critical condition (standard or proposed system using FP or SP). Experimental materials are available at the thesis repository (Dall, 2017).

10.2.2 Results and Discussion

Two participants were excluded from analysis as they were clearly not native English speakers and suitable replacement participants were found. Of the remaining 2929 responses, 307 were null responses (where the participant did not press the button) and excluded. As in the earlier experiment in Chapter 4 outliers were detected using 2.5 times the MAD (median absolute deviation) over all critical stimuli (Median=586, MAD=277.1). 206 responses were outliers and removed from the analysis. This left a final set of 2416 responses. Table 10.2 shows the mean, SD and N for each condition after outlier removal. For each voice the FP and SP condition were compared, and the FP and SP conditions across voices also. After applying Bonferroni correction, it was found that for the proposed voice RTs were significantly (t(1220)=2.171, $p<0.05$) faster for the SP condition than the FP condition (by 25.5ms on average). For the standard voice the SP condition RTs were also significantly (t(1192)=7.117, $p<0.0005$) faster than for the FP condition (by 67.7ms on average). That is, for both voices we again find the *opposite* effect of FPs compared to the natural results from Chapter 4 – namely slower RTs to a target word following an FP compared to an SP. Comparing across voices, there was no significant difference between the standard and proposed SP conditions (t(1230)=1.472, p=0.565), however, the difference between the proposed and standard FP conditions was significant (t(1182)=2.953, $p<0.05$). The RT in response to an FP for the proposed voice is 28.8ms faster than the standard voice.

That the RT to a target word is slower in response to synthetic FPs as compared to SP containing sentences is a replication of what was found in the earlier experiment in Chapter 4. It was the hope that a voice synthesising improved FPs as compared to a standard voice would be able to replicate effects found in natural speech. However, it *does* produce FPs which lead to a faster RT as compared to a standard voice. While

	Mean RT in ms	**SD in ms**	**N**
Proposed FP	615.5	162.9	595
Proposed SP	590.0	164.8	627
Standard FP	644.3	173.4	589
Standard SP	576.6	155.3	605

Table 10.2: Mean reaction times (RT) with standard deviations (SD) in ms for filled pause (FP) and silent pause (SP) conditions for the proposed and standard voice. N is *after* outlier removal.

the desired effect was not fully realised, it shows that the improvements suggested in this thesis have had not just a perceptual effect, but also a subconscious effect. The improvement is notable, but it is clear that the problem is not fully solved. Two things stands out when listening to the test stimuli. Firstly, seemingly one effect on FP realisation of combining the pronunciation variant alignment and synthesis techniques with the specific FP modelling, data mixing and proposed linguistic feature set is that many of the FPs are now preceded by some amount of pausing (e.g., EN2009cA3+4 at the thesis repository for the experiment (Dall, 2017)). This is a positive development as that is also what has been found in various corpus studies (see Chapter 4). However, in many instances a crackling noise filled those pauses (e.g., V1d119 at the thesis repository), something which is likely to have thrown participants off. Furthermore, the overall prosody of the speech is still far from being natural – this is probably the key issue – often the FPs sound out of place and this highlights that the prosodic structure around FPs is still not properly modelled, increased duration and lower F0 is only part of the equation (but a part that has an effect). One potential explanation is that the use of the mark switching technique, while improving acoustic base quality while synthesising more appropriate FPs, still primarily produces prosody based on read speech and that a prosodic discontinuity is introduced when switching from read to spontaneous speech as the models basis when synthesising the FP. It is likely that this problem needs to be solved before FPs can reproduce the natural speech effect.

10.3 A Final Evaluation

At the same time as the previous experiment a preference test was also performed using the standard and proposed voice (See Table 10.1). This was to evaluate the overall

	Standard	**Proposed**	**No Preference**
All	42.0%	**45.5%**	12.5%
FP	43.7%	**44.9%**	11.4%
No FP	40.3%	**46.1%**	13.6%

Table 10.3: Preference scores between the standard voice and the proposed voice.

effect of combining all proposed improvements in one voice and compare it to the standard voice. The standard voice and proposed voice from the previous experiment was used. 40 sentences were synthesised using each voice, 20 sentences containing no FPs and 20 sentences of FP containing sentences. The 20 sentences without FPs came from the set of 50 parallel sentences used throughout this thesis and the 20 sentences containing FPs were a subset of the 79 critical stimuli from the previous experiment. 38 paid native English speakers were recruited to take part. Each participant rated all 40 pairs of sentences and were asked to "pick which of the two sentences you think sounds more natural". The choices were "Sample 1", "Sample 2" and "No Preference". Experimental materials available at the thesis repository (Dall, 2017).

10.3.1 Results and Discussion

As previously, two participants were excluded from analysis due to clearly not being native speakers. Also, due to experimenter error, one participants' results were mixed up with a set of test results invalidating those results and was thus also excluded from analysis. This left a total of 1400 ratings from 35 participants. Table 10.3 shows the results. All "No Preference" ratings were divided evenly between systems to allow for significance testing using the exact binomial test. The overall difference between the two systems was not significant (p = 0.199). Looking at each type of sentences individually we can see that the difference between the systems is less for the FP containing sentences (p = 0.791) and although the preference was greater for the sentences not containing FPs this difference was also not significant (p = 0.130).

These results are disappointing and unexpected. It was expected that combining all the improvements, as documented throughout this thesis, would result in synthesis which is clearly perceptually preferred over a standard synthesis system's output. This should have been especially true for the FP containing sentences as a lot of attention has been paid to the synthesis of these phenomena. Instead, we see another situation, although the tendency is indeed toward preferring the proposed system, in which participants have no preference for either – in particular with regards to the FP containing sentences. Exactly what has caused this is not entirely clear. For the FP containing sentences the additional pause before many FPs noted in the previous experiment, combined with the odd scratching sounds also noted in the previous experiment, could have resulted in several of the FP containing sentences being clearly dispreferred (e.g., sentence V2d30 and V2b67 for the experiment at the thesis repository).

Table 10.4 details the duration and F0 of the synthesised sentences from the Standard and Proposed system. From this we can see that the proposed system does indeed produce FPs much more like those found in the natural data, both in terms of F0 fall and durations, although the general F0 around the FPs tend to be much higher than the natural ones. One notable thing is that the 'UH's of the proposed system are very long compared to the natural 'UH's, and this is quite likely a consequence of these additional pauses appearing (as they seem to have become part of the FP model itself instead of a general silence model), supporting the idea that these additional pauses can have had a disruptive effect. Note how the silence actually inserted by the synthesis system are generally much faster than the natural silences.

Another element could be the fact that in the previous experiments using FPs the existence of FPs was made explicit to the listeners as the FPs were in all the stimuli but in this case it was only in half the stimuli. It could very well be that since the FPs produced by a standard voice are often very short and almost unnoticeable (see Table 10.4), some participants may not have noticed them and thus preferred the sentence "without" an FP (as participants overwhelmingly did in Chapter 9). This could be further compounded by the difference in instructions – for the tests in Chapter 7 participants were purposely made aware of the presence of FPs (in the first test even told that they were of particular interest) and this could have affected peoples' ratings as found in Chapter 3. Thus the current test may not have tested what was intended. The explicit FP mention was not done in the test of linguistic features in Chapter 8, however, in that chapter the same FP representation was used, /UHV/, together with the same mark switching technique, such that the FP realisation only differed due to parsing. This realisation was found to be fairly similar, while the improvement was suggested to be from the linguistic feature set. An explanation for why the parsing might not have worked could lie in the fact that these sentences contained other conversational phenomena such as many repetitions, these phenomena might have had an adverse effect on the parsing accuracy to the degree in which these sentences were affected negatively.

	Dur mean (ms)	Dur median	F0 mean (Hz)	F0 median
Standard UH				
UH	90.2	85	182.5	179.2
Left Syll Vowel	81.0	75	185.5 (1.8)	176.3
Right Syll Vowel	78.0	75	185.6 (4.3)	183.2
Left Silence	70.0	70	-	-
Right Silence	70.0	70	-	-
Standard UM				
UM	165.1	140	190.8	184.9
Left Syll Vowel	-	-	196.5 (3.2)	183.0
Right Syll Vowel	-	-	196.8 (5.6)	189.0
Left Silence	70.3	70	-	-
Right Silence	286.3	70	-	-
Proposed UH				
UH	382.6	430	167.3	175.7
Left Syll Vowel	71.3	50	187.5 (20.3)	180.7
Right Syll Vowel	74.8	75	189.6 (20.4)	188.5
Left Silence	69.5	75	-	-
Right Silence	85	85	-	-
Proposed UM				
UM	368.9	390	173.6	167.7
Left Syll Vowel	-	-	198.4 (25.8)	186.7
Right Syll Vowel	-	-	186.8 (16.3)	180.4
Left Silence	62.6	60	-	-
Right Silence	277.5	85	-	-
Natural UH				
UH	221.1	214	174.5	175.1
Left Syll Vowel	99.4	70	177.8 (9.3)	180.0
Right Syll Vowel	67.6	60	180.0 (12.8)	178.5
Left Silence	205.4	90	-	-
Right Silence	211.1	158	-	-
Natural UM				
UM	373.8	360	170.3	167.5
Left Syll Vowel	-	-	184.8 (23.7)	181.8
Right Syll Vowel	-	-	170.1 (23.2)	173.4
Left Silence	199.8	90	-	-
Right Silence	228.4	190	-	-

Table 10.4: Details of phone durations and F0 for the read and spontaneous synthesis in the final test for ‘UH’ and ‘UM’ and their contexts. The number in parenthesis for left and right contexts is the mean deviation from the FP in context.

The non-FP containing sentences tended to be slightly more preferred, however again not significantly, and the reason for this is also unclear. Parsing and pronunciation reduction should both have contributed to a voice preferred over the standard, as shown by both methods previously in this thesis. It is possible this is due to the low number of sentences, when focusing on the No FP sentences only, giving little statistical power compared to earlier tests. Another potential explanation could lie in some previously unseen interaction between the use of categorical position representation and parsing combined with pronunciation variant alignment and fully reduced pronunciations. These methods have not been combined before in this thesis and perhaps using the more simplistic categorical position representation or parsing relies on a very stable pronunciation choice across training and synthesis in order to gather good statistics – and although fully reducing is closer to keeping consistency across training and synthesis than not reducing – it is still not fully consistent and could cause problems. A potential way of finding if this is the case would be to train a voice using full reduction for both alignment and synthesis and then apply the position representation and parsing techniques proposed in this thesis. This however, is considered out of scope of this thesis and is left as future work.

10.4 (Si)mply a (Re)search Front-end

A number of improvements have been suggested in this thesis – most of which touch on the front-end processing of the TTS system. Forced alignment, pronunciation choice, linguistic features and data mixing all rely on modifications to the front-end. As part of this work a tool has been written which can do all of these modifications. This tool is hereby released, as part of this thesis, as SiRe, (Si)mply a (Re)search Front-end, at *https://github.com/RasmusD/SiRe* under the Apache 2.0 license. It is written in Python 2.x and is designed with rapid linguistic feature experimentation in mind. I will here give an overview of features and design decisions. Together with this tool the corpora of read and spontaneous speech, and also the parallel 50 sentences, is released under the same license and can be found at the conference repository (Dall, 2017).

The utterance structure is fairly simple. At heart, an utterance is an object containing a list of words, which is an object containing a list of syllables, which is an object containing a list of phones. Each knows about its parent and its children. An utterance object has no other information stored than the name of the utterance itself. A word stores only its name and if parsing is applied also links to its father, grandfather and greatgrandfather phrase in the parse tree. A syllable its containing phones and a phone its identity and nothing else. Each, however, has methods to derive contextual features about itself such as a method to find its neighbouring phones, syllables, words and the like. This is a simple yet powerful structure as it does not assume any inherent relation between elements except a parent/sibling relationship. Not encoding relations to surrounding segments furthermore allows for insertion and deletion of additional segments through e.g. postprocessing rules without breaking the structural relations as these will be dynamically updated. An utterance can be created either from a monophone mlf from the output of the forced alignment procedure which contains the relevant word and syllable boundary markers and also syllable stress markers, but also from text. When creating an utterance from text (one line, one utterance), only words in the dictionary are available (Combilex and CMUdict are supported, though CMUdict must be run through the syllabifyCMUdict.py script to be compatible) for use. Any OOV words must be manually added.[1] Furthermore, no text normalisation except lower casing is performed, one must prenormalise any text desired to be uttered.[2] Finally, one can utilise phoneme reductions if present in the dictionary (present in Combilex not CMUdict) and either do standard selection, full reduction selection or a variable selection if n-gram scores for either the word or phone level is calculated using e.g. SRILM (Stolcke et al., 2011). When a pronunciation has been chosen it is checked against an accepted set of phones as defined by the dictionary in use – this set also defines the features of each phone used for context feature extraction.

The meat of the system is then the context feature extraction. This works by defining a "context skeleton" which lists all the appropriate context features which should be derived by a particular context set. In the skeleton each context is defined as being of either a boolean, integer or floating point type which restricts the allowed types of values stored in each context type (and these are also the supported feature types). They do not directly correspond to standard boolean, integers or floating points in terms of input type, rather they define what type of questions should be asked of the feature. A boolean feature is thus one of which Yes/No type questions should be asked – e.g., "Is this the /@/ phone?" – and the feature values can be of any type for which the == operator is defined (even objects), generally however this will be strings or actual

[1] One could easily add a G2P module, however, this has been here purposefully avoided. Phonetisaurus (*https://github.com/AdolfVonKleist/Phonetisaurus*) or Sequitur (*http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html*) models would be a decent choice.

[2] Again this is a conscious choice as the methods applied in this thesis does not touch upon text normalisation this was considered an unnecessary element to implement a system for. If one wish, one could incorporate e.g. Sparrowhawk (*https://github.com/google/sparrowhawk*) into SiRe.

booleans. An integer feature is a feature about which less/more than questions should be asked but which does not have any theoretical upper bound on its value (i.e., it cannot be immediately normalised to a value between 0 and 1),[3] such as "Is the absolute forward position of this phone in the utterance less than 7?" and the feature values *must* be of type int. A special out-of-utterance marker, "xx", is also accepted. Finally floating point types are features which should take a value between 0 and 1 and of which less/more than questions should also be asked – this is distinct from the integer in some DNN implementations in which integer types will be binarised but floating types stay of type floating point. Feature values must be of either a floating point between 0 and 1 or an integer or floating point with a defined range of values which can be converted into a value between 0 and 1.[4] A special out-of-utterance marker, "xx", is also accepted and the value of 0 is reserved for this for DNNs as they would else need a binary marker for this.[5]

Each context skeleton is a class which can inherit all features from another class by utilising it as a base class. This allows for rapid and easy definition and redifinition of a feature set as features can be shared and overriden as any other variables in a class. With each context skeleton goes a method to extract and add all relevant contexts from an utterance into the skeleton. This method has a soft limitation on it which means that it cannot add features not in the skeleton, however it *can* not add a feature present in the skeleton; in this case a warning will be issue. Once a set of features have been derived from a corpus of text or a monophone mlf with phone durations, full context labels and question sets based on the derived features can be created. Labels and question sets suitable for the following three systems can be created – HTS HMM, HTS DNN and Merlin[6]. Each full context labels will contain all features in a given feature set, of which sets using categorical, absolute and relational positional values and also PCFG and dependency parsing are available in all thinkable combinations besides the standard feature set.[7]

[3] Theoretically no lower bound either, however, no features in any of the described sets have meaningful negative values.

[4] For HTS, internally all values are converted to an integer between 0 and 100 due to HTS treating a dot in a context string as a wildcard. The output for an HTS style decision tree will thus actually be a value between 0 and 100, whereas for a DNN floating point input node the true floating point is still used.

[5] With the consequence that something else which should have been 0.00 will instead be given the 0.01 value, and a slight shift in meaning for e.g. relational values.

[6] A recent NN TTS back-end from the CSTR (*https://github.com/CSTR-Edinburgh/merlin*).

[7] If proper POS tagging is wanted for the standard feature set this can be obtained through the PFCG parse tree without adding all of the other PCFG features. But then, if you've already parsed it why not use the better set?

In addition to these, the system can output monophone mlfs and slf lattices suitable for both standard and pronunciation variant forced alignment.[8] A number of other utility scripts exist to support the creation and modification of parse trees and other data for use with SiRe, simple support for using and outputting state aligned mlfs and labels (for use in NN systems such as Merlin) and also a few tools for corpus analysis and OOV word detection.

All in all, this yields a tool which hopefully will lower the bar for other researchers interested in the linguistic context feature set as compared to currently available front-ends such as Festival.

10.5 Chapter Conclusions

In this chapter two final evaluations of the methods developed in this thesis has been presented. The RT experiment from 4 was re-run to evaluate the effect of the proposed modifications to FP synthesis and TTS voice creation. It was found that although synthetic FPs still produce slower RTs to a target word than SPs for the proposed voice – the proposed voice did indeed improve FP synthesis compared to a standard voice as it allowed for faster RTs than that. In the second evaluation a preference test was run between the proposed and standard voice using both FP containing and sentences without FPs. It was found that although there was a tendency toward prefering the proposed voice, this was not a statistically significant effect. It was proposed that the reason for this result is to be found partly in participants dis-preference for clearly disfluent speech as being natural, partly in the instructions used, partly due to an issue with FP realisation noted in both experiments and partly due to some interaction between the pronunciation variant methods and the new linguistic feature set – as these methods had not been combined before.

Finally a research front-end, SiRe, has been presented and released for all uses which allows for replication of the proposed methods and, it is hoped, for other researchers to pick up and extend the work presented in this thesis.

[8] Please note that these may differ from those presented in this thesis as a separate non-releasable system, by Korin Richmond, was initially used for this.

Chapter 11

Conclusions and Future Directions for Conversational Speech Synthesis

In this thesis a number of hypothesis have been put forth and tested in various manners. The central hypothesis of this thesis was that:

> **Main Hypothesis:** The use of spontaneous conversational data and phenomena can lead to improvements in the perception of TTS output.

This was tested through a number of sub-hypothesis which will here be evaluated.

> **Hypothesis 1:** Spontaneously produced natural conversational speech is considered more natural than read aloud prompts.

In Chapter 3, this claim was tested directly by evaluating the naturalness of naturally produced read and spontaneous speech and its faithful transcriptions. It was shown that conversational speech was rated as more natural, in a mean opinion score test, than their read counterparts – even when instructions were biased towards reading naturalness. This tendency persisted even after controlling for speaker and content. It was, however, found that if considering only differences in content (i.e., rating text only) then read text was considered more natural in the reading naturalness instructions (but not in standard or conversational naturalness). This strongly supports Hypothesis 1.

> **Hypothesis 1a:** TTS voices can benefit from the use of spontaneous speech in the training corpus.

This claim was tested in a number of ways throughout the thesis. In Chapter 3 an initial investigation comparing read and spontaneous speech-based HMM TTS voices it was shown that read speech-based voices are, using current standard techniques, preferred over voices based on spontaneous speech and that this preference persist even when including spontaneous speech phenomena such as filled pauses, discourse markers and repetitions. Sub-par modelling of the spontaneous speech was hypothesised to be the problem and in Chapters 5 and 6 two investigations into improving the modelling of spontaneous speech was presented. The first investigation, pronunciation variant forced alignment, aimed to improve phone segmentation of the speech by allowing for pronunciation variants in the procedure. This was found to improve spontaneous speech synthesis, however for a read speech-based voice it resulted in hyperarticulated speech. The second method, pronunciation variant synthesis, focused on producing pronunciations more akin to those found in spontaneous speech. Not being variant, but simply producing reduced pronunciations when possible was found to improve both read and spontaneous speech-based synthesis – however, it was also concluded that directly modelling spontaneous speech would not yield a sufficiently high quality of speech to allow investigating filled pause synthesis. In Chapter 7, a number of methods for mixing read and spontaneous speech for voice modelling was investigated with the aim of retaining the higher base quality of the read speech while still benefiting from the presence of filled pauses in the spontaneous speech corpus. A corpus marking technique in which the mark was switched from read to spontaneous for the filled pause while also using a specific filled pause phone model was found to improve filled pause synthesis. Overall, Hypothesis 1a has not been directly vindicated, however, it has been shown that for phenomena such as filled pauses for which no training data exist in standard corpora, the inclusion of spontaneous speech data can be beneficial.

It is worth noting here that the used corpus of spontaneous speech contained approximately 1h of speech. This, as mentioned in Chapter 3 is just enough for reasonable synthesis. However, it is quite possible that some of the results found has been affected by the small dataset size and that increasing the amount of spontaneous data is necessary in order to produce good quality synthesis from spontaneous speech. This seems to be fairly likely given the increased variability of spontaneous speech, and this variability may necessitate more data for robust modelling. Spontaneous speech data, transcribed and suitable for TTS is unfortunately a very laborious and expensive process, inherently limiting the amount of data which is possible to collect, perhaps in future work this can be remedied and tested.

Hypothesis 1b: TTS voices can benefit from being more spontaneous –

whether based on read or spontaneous speech.

This claim was in part investigated through the various alternative pronunciation methods in Chapter 6, but also through the filled pause synthesis in Chapter 7 and linguistic feature sets in Chapter 8. Using more reduced pronunciations is more akin to natural spontaneous speech (as determined by the investigation in Chapter 5) and the investigation confirmed that this creates more natural sounding speech. Filled pauses synthesis was improved in Chapter 7 and the perception of the speech also improved. However, as shown in Chapter 9 and suggested in Chapter 10, synthetic speech containing filled pauses is not rated favourably compared to speech without filled pauses, even with better filled pause modelling. The improvements from parsing and position representation in Chapter 8 is suggested to come mainly from better phrase level prosodic modelling, and while this phrase level prosody in a read voice is read speech like, it still has a decidedly conversational element to it. Overall, some benefit from producing more spontaneous speech has been found in this thesis, however, the potential benefits found for natural speech in Chapter 3 have not been fully realised – Hypothesis 1b has thus to some extend been vindicated.

> **Hypothesis 2:** Pronunciation variation is important both to model and to realise conversational speech.

Chapters 5 and 6 tested this directly by introducing pronunciation variance into the forced alignment procedure and at synthesis time. Both methods improved the perception of TTS voices based on spontaneous speech, however at synthesis time simply producing reduced output with no predictive variance was as good as using word-level n-gram language models to predict which words to reduce. In general, the findings support Hypothesis 2.

> **Hypothesis 2a:** Standard read speech-based voices can also benefit from this.

A read speech-based voice did not improve in perceived naturalness when applying pronunciation variant forced alignment. This, however, seemed to be due to hyperarticulation as a result of a cleaner underlying acoustic model, and the following investigation into pronunciation choice shows that a read speech voice can benefit from reduced pronunciation choice more akin to that found in spontaneous speech. Furthermore the potential is particularly great in a voice based on pronunciation variant forced alignment as compared to standard alignment due to its cleaner underlying acoustic model. Hypothesis 2a is thus supported by these findings.

> **Hypothesis 3:** Filled pauses can provide benefits to the listener in TTS.

The psycholinguistic experiments performed in Chapter 4 highlights some potential subconscious benefits of filled pauses to the listener – namely faster reaction times to a target word and higher change detection rate. It was shown that in standard synthetic speech neither effect appeared, and that, in fact, FPs seemed to have an adverse effect in terms of reaction time as compared to a silent pause. The reaction time effect appeared when using vocoded speech, illustrating that it should be possible for a synthetic voice to replicate the effect, however for the change detection experiment the effect did not appear for vocoded speech – illustrating limits to current synthesis ability to replicate filled pause benefits. This suggests that, with current vocoding techniques, the potential benefits of filled pauses are not universally translatable into synthesis. Furthermore the ability to evaluate filled pause impact on TTS is complicated by the finding in Chapter 9 that listeners prefer the same sentence without any FPs as compared to FPs being present, and this must be kept in mind when performing such an evaluation as suggested by the results in Chapter 10.

> **Hypothesis 3a:** Filled pauses can be usefully employed in a TTS system if properly realised.

In Chapter 7 methods for improved filled pause synthesis were proposed and it was shown that by using a specific phone model, a corpus marking technique combined with a mark switching method at synthesis time significantly improved both the replication of acoustic effects of FPs noted in Chapter 4 and the perception of FP containing synthesis. In Chapter 8 this was further improved by the addition of positional and parsing-based modifications to the linguistic feature set. This did not result in significantly improved perception of filled pause containing sentences when combined with the pronunciation variant methods of Chapters 5 and 6 in the test in Chapter 10 – notably however, this test was not solely focused on FPs and the finding of Chapter 9 suggests this result to be expected in such a case, whether or not the filled pauses are realised any better. Overall, Hypothesis 3a has been neither vindicated nor refuted, the results show that FPs, in their current realisation by the proposed system, are not a benefit to the TTS system. However, the results of the reaction time experiment re-run in Chapter 10, and the evaluations in Chapters 7 and 8, shows that some improvement has been made in the effect of filled pauses on listeners and the hope is that further improvements could vindicate Hypothesis 3a.

> **Hypothesis 3b:** Filled pause insertion can be accurately predicted from text.

This claim is directly tested in Chapter 9 in which several methods for filled pause insertion prediction was presented. It was shown that there is a pattern to both human usage and human prediction of filled pauses usage and that this pattern can be predicted by a combination of recurrent neural network and n-gram language modelling. This is further extended to prediction of multiple filled pauses and multiple filled pause types in a lattice-based framework which again achieved very high prediction accuracy when compared to human performance. This supports Hypothesis 3b.

Overall the Main Hypothesis has been tested in a variety of different ways and several of these support the main tenet of this hypothesis. Namely that spontaneous conversational speech and phenomena can improve TTS perception. While using spontaneous speech data directly, even after improved modelling, did not provide a measurable benefit over read speech-based voices – read speech-based voices did improve after application of spontaneous speech inspired methods, such as pronunciation reduction and linguistic feature analysis, showing the potential in utilising methods inspired by spontaneous speech phenomena. Furthermore the initial result of the perceptual tests in Chapter 3 promises further future improvements should spontaneous speech-based TTS be further pursued.

11.1 Contributions

This thesis has contributed to the general body of knowledge, tools and methods in the following ways:

- It has been shown that spontaneously produced speech is considered more natural than read aloud prompts (Chapter 3).
- A corpus of parallel spontaneous and read speech from the same female British English speaker has been developed and released for use by other researchers.
- The potential subconscious benefits of using filled pauses has been shown in a reaction time and a change detection paradigm (Chapter 4).
- Current vocoding and synthesis techniques has been tested in the above mentioned experimental paradigms and shown to fall short.

- A new effect for natural speech, the influence of speaking rate on reaction time (faster speaking rate = faster reactions), has been found.

- It has been shown that current alignment techniques produce more, and more serious, errors on spontaneous speech than on read speech (Chapter 5).

- Improved spontaneous speech modelling has been developed though both pronunciation variant forced alignment and pronunciation choice (Chapters 5 and 6.

- The pronunciation variant forced alignment method also improve the underlying acoustic model of a read speech-based voice, allowing it to benefit from improved pronunciation choice.

- Improved pronunciation choice, inspired by spontaneous speech effects, also improve read speech-based voices using a standard alignment method (Chapter 6).

- Through these methods it has also been shown that consistency across training and synthesis provides clear benefits. However, this consistency can be beneficially broken if better modelling is applied. Thus ideally one would stay consistent while also doing better modelling.

- The synthesis of filled pauses has been improved through a specific phone model (Chapter 7).

- The synthesis of filled pauses has been improved through a data mixing technique utilising corpus marking during training and mark switching for filled pauses at synthesis time.

- These techniques allow for utilising spontaneous speech data for phenomena not present in read speech without degrading overall quality.

- General read speech synthesis quality has been improved through position and parsing (Chapter 8).

- Using a categorical positional representation improves TTS perception and decreases model complexity.

- Probabilistic Context Free Grammar parsing improves TTS perception.

- Combining PCFG parsing with dependency parsing provides further benefit.

- These methods also improve perception of synthesis of sentences containing filled pauses when combined with the specific phone modelling and a data mixing technique.

- Humans are consistent in their usage of filled pauses and so this usage should be predictable (Chapter 9).

- Humans filled pause prediction is consistent with actual usage so human accuracy can be used as a benchmark for automatic methods.

- Filled pauses insertion position matters perceptually, however, if obvious to human listeners are not preferred over not being used.

- A combination of n-gram and recurrent neural network language models can be used to insert multiple filled pause types into multiple insertion points supported by a general disfluency parameter and the performance is close to human.

- The here developed improvements to TTS in general and filled pauses synthesis in particular can improve the reaction time of participants as compared to using a standard TTS voice (Chapter 10).

- A research front-end, SiRe, which can be used to produce any of the proposed methods of this thesis has been developed and released for free use by the wider TTS community.

11.2 Future Work

While this thesis has provided many contributions there is also a number of potential avenues for future research which have either not been addressed or opened up as a result of the work done in this thesis. Here is a non-exhaustive list of some of the more pertinent, in the opinion of the current author, questions:

- Could a professional voice talent with suitable training produce convincing filled pauses from prompts removing the need for spontaneous speech for filled pause synthesis?

- Could acted conversational speech, such as that from playwriting, provide an intermediate type of “spontaneous” speech, which is easier to model but retains defining characteristics?

- Does the additional variability of spontaneous speech make it such that more data is needed for good modelling? I.e., is 1 hour of speech data sufficient even if better modelling techniques are applied?

- What is it about vocoding that limits its ability to reproduce the change detection effect? Perhaps investigations into cognitive load of the listener could shed some light on this?

- Specifically, it is proposed that, by running a dual attention task with change detection as one part on natural speech one could see if the effect still appears. If not, this would support an increased cognitive load of vocoded speech hypothesis.

- Improved pronunciation choice has a clear potential benefit and more sophisticated models for the creation of potential pronunciation variants should be investigated. If these methods can be deployed consistently across training and synthesis so much the better.

- The usage of filled pauses and more conversational TTS should be evaluated in a more ecologically valid scenario than single sentences played in sound-proofed booths.

- Specifically, scenarios involving virtual avatars which are either spontaneous or hesitant would provide useful real-world scenarios.

- Filled pause insertion methods could also be usefully evaluated in similar scenarios.

- Sentence level prosody would need to be further improved for filled pauses. Template-based approaches are one possibility.

- The linguistic context feature set should be further investigated.

- Other types of parsing, other combinations with positional representations. Word embeddings, utterance level and dialogue state features are further possibilities.

- Could one leverage paraphrasing techniques to rewrite input text to be more conversational in nature, achieving more conversational TTS not through traditional TTS methods but directly through discourse?
- Robust methods for accurately modelling the increased variability in spontaneous speech should be developed.
- These methods could take many forms, such as improved model clustering, increased data sizes or different acoustic models (e.g., NN's – see below).
- A tighter coupling of the TTS system and a potential dialogue system could provide dialogue level context useful for varying the speech.
- Some streams (F0, MCEP, BAP, durations) may be better modelled by one type of speech over the other, e.g. using spontaneous intonation and speed but read spectral features.
- Streams are currently independent on each other but could perhaps be put in a hierarchical order such as the output of one stream could be used as the input to another. E.g., there is likely an interplay between duration and F0.
- F0 modelling was an issue with the spontaneous speech not dealt with in this thesis. However, while working with the data, it became clear that many F0 modelling techniques deal less well with spontaneous speech as compared to read speech. Thus investigations into more robust F0 modelling seems desirable.
- The focus of this thesis has been English, but FPs are utilised in most other languages and the findings here could be usefully evaluated/replicated in other languages. Some work has been done for RT experiments in Dutch, but not in the context of synthesis. Even creating an overview of FPs in a number of different languages would be useful.

This list is of course non-exhaustive, but hopefully will provide the interested reader with suggestions for future directions should one wish to work with conversational speech.

11.3 Final Remarks

This thesis has been concerned with primarily HMM-based TTS. During the work for this thesis another major method for SPSS has appeared – NN-based TTS. Little mention has been made of this, nor unit selection synthesis (the third major current method). While many of the methods and ideas here presented translate directly into use for these methods I will end this thesis with a few words on how either of these TTS paradigms may work differently with the methods of this thesis.[1]

In terms of unit selection a few things are primarily worth mentioning. As in unit selection we are doing re-shuffling and playback of actual samples of natural speech the method has great potential for replicating the psycholinguistic results found for natural speech. The main challenge would be to identify and use the correct units from a database containing both read and spontaneous speech, but once this is done well the effects should be more likely to appear, particularly the change detection effect is suddenly plausibly replicated. Elsewise improved word segmentation, pronunciation and linguistic feature sets should all also improve a unit selection voice – and filled pause prediction from text is unaffected. However, the datamixing methods, are less likely to provide as good results as finding suitable units crossing the boundary between the two types of speech is potentially hard – one solution could be a hybrid system in which the unit selection is driven by a parametric system (effectively the best performing systems in recent years).

For neural network synthesis, the methods here presented are essentially unaffected. One interesting feature of neural network synthesis is the replacement of the decision tree context clustering by the net. This allows for some interesting methods not possible in an HMM system. The datamarking method relies on a binary mark – read or spontaneous. This would always be one or the other at training time but at synthesis time, as the input node representing this distinction is actually a floating point, one can set this to any value in between – potentially interpolating naturally between the two modes in the network. This could also have positive effects on duration modelling by allowing for durations closer to that found in spontaneous speech without fully committing to the much higher spontaneous speaking rate. In fact, the interested reader should read the thesis of Emily Dreke (Dreke, 2016), co-supervised by the present author, which touches upon this and also on evaluation in an avatar scenario.

In the end it is the belief of the current author that research into more conversational style TTS will be a fruitful avenue of research for many years and hopes that this thesis represents a useful contribution to the debate.

[1] Thus if a method is not mentioned it is assumed it will behave as found for HMM synthesis.

Bibliography

Adell, J., Bonafonte, A., and Escudero, D. (2006). Disfluent Speech Analysis and Synthesis: a preliminary approach. In *Proc. Speech Prosody*, Dresden, Germany.

Adell, J., Bonafonte, A., and Escudero, D. (2007a). Filled Pauses in Speech Synthesis: Towards Conversational Speech. In *Proc. TCD*, pages 358–365, Pilsen, Chech Republic.

Adell, J., Bonafonte, A., and Escudero, D. (2007b). Statistical Analysis of Filled Pauses Rhytm for Disfluent Speech Synthesis. In *Proc. SSW*, pages 223–227, Bonn, Germany.

Adell, J., Bonafonte, A., and Escudero, D. (2010a). Modelling Filled Pauses Prosody to Synthesise Disfluent Speech. In *Proc. Speech Prosody*, Chicago, USA.

Adell, J., Bonafonte, A., and Escudero, D. (2010b). Synthesis of Filled Pauses Based on a Disfluent Speech Model. In *Proc. ICASSP*, Dallas, Texas, USA.

Adell, J., Escudero, D., and Bonafonte, A. (2012). Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, 54(3):459–476.

Agiomyrgiannakis, Y. (2015). Vocaine the vocoder and applications in speech synthesis. In *Proc. ICASSP*, pages 4230–4234, Brisbane, Australia.

Andersson, S. (2013). *Synthesis and Evaluation of Conversational Characteristics in Speech Synthesis*. PhD thesis, University of Edinburgh.

Andersson, S., Georgila, K., Traum, D., Aylett, M., and Clark, R. A. J. (2010a). Prediction and Realisation of Conversational Characteristics by Utilising Spontaneous Speech for Unit Selection. In *Proc. Speech Prosody*.

Andersson, S., Yamagishi, J., and Clark, R. (2010b). Utilising Spontaneous Conversational Speech in HMM-Based Speech Synthesis. In *Proc. SSW*, Kyoto, Japan.

Andersson, S., Yamagishi, J., and Clark, R. A. (2012). Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis. *Speech Communication*, 54(2):175–188.

Andreas, J., Klein, D., and Division, C. S. (2015). Alignment-Based Compositional Semantics for Instruction Following. In *Proc. EMNLP*, pages 1165–1174, Lisbon, Portugal.

Aylett, M. P., Dall, R., Ghoshal, A., Henter, G. E., and Merritt, T. (2014). A Flexible Front-End for HTS. In *Proc. Interspeech*, Singapore.

Barra-Chicote, R., Yamagishi, J., Montero, J. M., Watts, O., King, S., and Macias-Guarasa, J. (2010). The GTH-CSTR Entries for the Speech Synthesis Albayzin 2010 Evaluation: HMM-based Speech Synthesis Systems considering morphosyntactic features and Speaker Adaptation Techniques. In *Proc. II Iberian SLTech Workshop*, pages 353–358.

Barzilay, R. and McKeown, K. R. (2001). Extracting Paraphrases From a Parallel Corpus. In *Proc. ACL*, pages 50–57, Morristown, NJ, USA.

Baumann, T. (2013). *Incremental Spoken Dialogue Processing: Architecture and Lower-level Components*. PhD thesis, Universität Bielefeld, Germany.

Baumann, T. and Schlangen, D. (2012). INPRO_iSS: A Component for Just-In-Time Incremental Speech Synthesis. In *Proc. ACL System Demonstrations*, Jeju Island, South Korea.

Baumann, T. and Schlangen, D. (2013). Open-ended, Extensible System Utterances Are Preferred, Even If They Require Filled Pauses. In *Proc. SIGdial*, Metz, France.

Bechet, F., Nasr, A., and Favre, B. (2014). Adapting dependency parsing to spontaneous speech for open domain spoken language understanding. In *Proc. Interspeech*, pages 135–139, Singapore.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, 113(2):1001.

Belz, A. (2005). Statistical Generation: Three Methods Compared and Evaluated. In *Proc. European Workshop on Natural Language Generation*, Aberdeen, Scotland.

Bennett, C. L. (2005). Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005. In *Proc. Interspeech*, pages 105–108, Lisbon, Portugal.

Benoit, C., Grice, M., and Hazan, V. (1996). The SUS test : A method for the assessment of text-to-speech synthesis intelligibillity using Semantically Unpredictable Sentences. *Speech Communication*, 18:381–392.

Binnenpoorte, D., Cucchiarini, C., Strik, H., and Boves, L. (2004). Improving automatic phonetic transcription of spontaneous speech through variant-based pronunciation variation modelling. In *Proc. LREC*, pages 681–684, Lisbon, Portugal.

Blaauw, E. (1991). Phonetic Characteristics of Spontaneous and Read-Aloud Speech. In *Proc. ESCA Workshop on Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication*, pages 1–5, Barcelona, Spain.

Blaauw, E. (1992). Phonetic differences between read and spontaneous speech. In *Proc. ICSLP*, pages 751–754, Banff, Canada.

Black, A. W. and Lenzo, K. A. (2001). Flite: a small fast run-time synthesis engine. In *Proc. SSW*, Perthshire, Scotland.

Black, A. W., Taylor, P., and Caley, R. (2014). Festival 2.4 Documentation.

Black, A. W. and Tokuda, K. (2005). The Blizzard Challenge 2005 : Evaluating corpus-based speech synthesis on common datasets. In *Proc. Interspeech*, pages 77–80, Lisbon, Portugal.

Blackmer, E. R. and Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39(3):173–94.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., and Brennan, S. E. (2001). Disfluency rates in conversation: effects of age, relationship, topic, role, and gender. *Language and speech*, 44(2):123–47.

Brennan, S. (2001). How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language*, 44(2):274–296.

Brognaux, S. (2015). *Expressive speech synthesis: Research and system design with hidden Markov models*. PhD thesis, Université catholique de Louvain and Université de Mons.

Brognaux, S. and Drugman, T. (2014). Phonetic variations: Impact of the communicative situation. In *Proc. Speech Prosody*, Dublin, Ireland.

Brognaux, S., Drugman, T., and Saerens, M. (2014a). Synthesizing sports commentaries: One or several emphatic stresses? In *Proc. Speech Prosody*, Dublin, Ireland.

Brognaux, S., Picart, B., and Drugman, T. (2013). A New Prosody Annotation Protocol for Live Sports Commentaries. In *Proc. Interspeech*, number August, Lyon, France.

Brognaux, S., Picart, B., Drugman, T., and Louvain, D. (2014b). Speech synthesis in various communicative situations: Impact of pronunciation variations. In *Proc. Interspeech*, Singapore, Singapore.

Bulyko, I., Ostendorf, M., Siu, M., Ng, T., Stolcke, A., and Çetin, Ö. (2007). Web resources for language modeling in conversational speech recognition. *ACM Transactions on Speech and Language Processing*, 5(1):1–25.

Burkhardt, F. and Sendlmeier, W. F. (2000). Verification of Acoustical Correlates of Emotional Speech using Formant Synthesis. In *Proc. ISCA Workshop on Speech and Emotion*, pages 151–156, Newcastle, Northern Ireland, UK.

Cadic, D. and Segalen, L. (2008). Paralinguistic Elements in Speech Synthesis. In *Proc. Interspeech*, pages 1861–1864, Brisbane, Australia.

Cakmak, H. (2016). *Audiovisual Laughter Synthesis - A Statistical Parametric Approach*. PhD thesis, Faculte Polytechnique de Mons.

Callison-Burch, C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. In *Proc. EMNLP*, page 196, Morristown, NJ, USA.

Campbell, N. (1998). Where is the information in speech? (And to what extend can it be modelled in synthesis?). In *Proc. ESCA/COCOSDA 3rd Speech Synthesis Workshop*, Jenolan Caves, Australia.

Campbell, N. (2004). Speech & Expression; the Value of a Longitudinal Corpus. In *Proc. LREC*, pages 183–186, Lisbon, Portugal.

Campbell, N. (2006a). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1171–1178.

Campbell, N. (2006b). On the Structure of Spoken Language. In *Proc. Speech Prosody*, Dresden, Germany.

Campbell, N. (2007). Towards Conversational Speech Synthesis; Lessons Learned from the Expressive Speech Processing Project. In *Proc. SSW*, pages 22–27, Bonn, Germany.

Campbell, W. N. (1997). Synthesizing spontaneous speech. In *Computing Prosody*, pages 165–186. Springer.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus*. *Language Resources and Evaluation Journal*, 41(2):181–190.

Charniak, E. and Johnson, M. (2001). Edit detection and parsing for transcribed speech. In *Proc. NAACL*, pages 118–126, Morristown, NJ, USA.

Chen, D. and Manning, C. D. (2014). A Fast and Accurate Dependency Parser using Neural Networks. In *Proc. EMNLP*, pages 740–750, Doha, Qatar.

Chen, X., Wang, Y., Liu, X., Gales, M. J. F., and Woodland, P. C. (2014). Efficient GPU-based training of recurrent neural network language models using spliced sentence bunch. In *Proc. Interspeech*, pages 641–645, Singapore.

Christenfeld, N. (1995). Does it hurt to say um? *Journal of Nonverbal Behavior*, 19(3):171–186.

Cieri, C., Miller, D., and Walker, K. (2004). The Fisher Corpus: A Resource for the Next Generations of Speech-to-Text Fisher. In *Proc. LREC*, Lisbon, Portugal.

Clark, H. H. and Fox, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84:73–111.

Clark, H. H. and Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive psychology*, 37(3):201–42.

Clark, R. A. J., Podsiadlo, M., Fraser, M., Mayo, C., and King, S. (2007). Statistical analysis of the Blizzard Challenge 2007 listening test results. In *Proc. Blizzard Challenge Workshop*, pages 1–6, Bonn, Germany.

Clark, R. A. J., Richmond, K., and King, S. (2004). Festival 2 Build Your Own General Purpose Unit Selection Speech Synthesiser. In *Proc. SSW*, Pittsburgh, USA.

Collard, P. (2009). *Disfluency and listeners' attention: An investigation of the immediate and lasting effects of hesitations in speech.* PhD thesis, University of Edinburgh.

Corley, M. and Hartsuiker, R. J. (2003). Hesitation in speech can. . . um. . . help a listener understand. In *Proc. 25th meeting of the Cognitive Science Society*, Boston, USA.

Corley, M. and Hartsuiker, R. J. (2011). Why um helps auditory word recognition: the temporal delay hypothesis. *PloS one*, 6(5):e19792.

Corley, M., MacGregor, L. J., and Donaldson, D. I. (2007). It's the way that you, er, say it: hesitations in speech affect language comprehension. *Cognition*, 105(3):658–68.

Dall, R. (2017). Thesis Material - Rasmus Dall, [dataset]. University of Edinburgh. Centre for Speech Technology Research. http://dx.doi.org/10.7488/ds/1996 .

Dall, R., Brognaux, S., Richmond, K., Valentini-botinhao, C., Henter, G. E., Hirschberg, J., Yamagishi, J., and King, S. (2016a). Testing the Consistency Assumption: Pronounciation Variant Forced Alignment in Read and Spontaneous Speech Synthesis. In *Proc. ICASSP*, Shanghai, China.

Dall, R., Hashimoto, K., Oura, K., Nankaku, Y., and Tokuda, K. (2016b). Redefining the Linguistic Context Feature Set for HMM and DNN TTS Through Position and Parsing. In *Proc. Interspeech*, San Francisco, CA, USA.

Dall, R., Tomalin, M., Wester, M., Byrne, W., and King, S. (2014a). Investigating Automatic & Human Filled Pause Insertion for Speech Synthesis. In *Proc. Interspeech*, Singapore.

Dall, R., Wester, M., and Corley, M. (2014b). The Effect of Filled Pauses and Speaking Rate on Speech Comprehension in Natural, Vocoded and Synthetic Speech. In *Proc. Interspeech*, Singapore.

Dall, R., Wester, M., and Corley, M. (2015). Disfluencies in Change Detection in Natural, Vocoded and Synthetic Speech. In *Proc. Disfluencies in Spontaneous Speech*, Edinburgh, Scotland, UK.

Dall, R., Yamagishi, J., and King, S. (2014c). Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation. In *Proc. Speech Prosody*, Dublin, Ireland.

de Marneffe, M.-C. and Manning, C. D. (2015). Stanford typed dependencies manual.

Dreke, E. (2016). *Conversational Speech Synthesis*. Msc thesis, The University of Edinburgh.

Durrett, G. and Klein, D. (2015). Neural CRF Parsing. In *Proc. ACL*, Beijing, China.

Eichner, M., Wemer, S., Wolfi, M., and Hoffmann, R. (2003). Towards spontaneous speech synthesis - lm based selection of pronunciation variants. In *Proc. ICASSP*, pages 248–251.

Eichner, M., Wolff, M., and Hoffmann, R. (2002). Improved Duration Control for Speech Synthesis Using a Multigram Language Model. In *Proc. ICSSP*, pages 417–420, Orlando, Florida, USA.

Eide, E., Aaron, A., Bakis, R., Hanza, W., Picheny, M., and Pirelli, J. (2004). A Corpus-Based Approach To <AHEM/> Expressive Speech Synthesis. In *Proc. ISSW*, pages 79–84, Pittsburgh, USA.

Fackrell, J., Skut, W., and Hammervold, K. (2003). Improving the accuracy of pronunciation prediction for unit selection TTS. In *Proc. Eurospeech*, pages 2473–2476, Geneva, Switzerland.

Fitzgerald, E. and Jelinek, F. (2008). Linguistic Resources for Reconstructing Spontaneous Speech Text. In *Proc. International Conference on Language Resources and Evaluation*, number Ldc, pages 3449–3452, Marrakech, Morocco.

Fitzgerald, E. C. (2009). *Reconstructing Spontaneous Speech*. PhD thesis, The Johns Hopkins Univeristy.

Fox Tree, J. E. (1995). The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language*, 34(6):709–738.

Fox Tree, J. E. (2000). Coordinating Spontaneous Talk. In Wheeldon, L. R., editor, *Aspects of Language Production*, pages 375–406. Psychology Press, Philadelphia.

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory and Cognition*, 29(2):320–326.

Fox Tree, J. E. and Schrock, J. C. (1999). Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes. *Journal of Memory and Language*, 40(2):280–295.

Fraser, M. and King, S. (2007). The Blizzard Challenge 2007. In *Proc. Blizzard Challenge Workshop*, pages 1–12.

Ganitkevitch, J., Callison-burch, C., Napoles, C., and Durme, B. V. (2011). Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation. In *Proc. EMNLP*, Edinburgh, Scotland.

Goddijn, S. and Binnenpoorte, D. (2003). Assessing Manually Corrected Broad Phonetic Transcriptions in the Spoken Dutch Corpus. In *Proc. ICPhS*, pages 1361–1364, Barcelona, Spain.

Goodfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings ICASSP*, pages 517–520, San Francisco, CA, USA.

Gotoh, Y. and Renals, S. (2000). Sentence Boundary Detection in Broadcast Speech Transcripts. In *Proc. ISCA Workshop: Automatic Speech Recognition: Challenges for the New Millennium*, number September, pages 228–235.

Gutkin, A., Gonzalvo, X., Breuer, S., and Taylor, P. (2010). Quantized HMMs for Low Footprint Text-To-Speech Synthesis. In *Proc. Interspeech*, Makuhari, Chiba, Japan.

Hashimoto, K., Oura, K., Nankaku, Y., and Tokuda, K. (2015). The effect of neural networks in statistical parametric speech synthesis. In *Proc. ICASSP*, pages 4455–4459, Brisbane, Australia.

Heeman, P. A. and Allen, J. F. (1999). Speech Repairs, Intonational Phrases, and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialogue. *Computational Linguistics*, 25(4):527–571.

Henter, G. E., Merritt, T., Shannon, M., Mayo, C., and King, S. (2014). Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech. In *Proc. Interspeech*, Lyon, France.

HTS, W. G. (2015). HTS Demo 2.3: Label Overview. http://hts.sp.nitech.ac.jp/archives/1.0/HTS-demo-CMU-Communicator.tar.gz.

Hu, Q., Richmond, K., Yamagishi, J., and Latorre, J. (2013). An experimental comparison of multiple vocoder types. In *Proc. SSW*, pages 135–140, Barcelona, Spain.

Huang, S. and Renals, S. (2010). Hierarchical Bayesian Language Models for Conversational Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):1941–1954.

ITU (2014). Method for the subjective assessment of intermediate quality level of audio systems. In *ITU Recommendation ITU-R BS.1534-2*. International Telecommunication Union Radiocommunication Assembly, Geneva, Switzerland.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2000). Probabilistic Relations between Words: Evidence from Reduction in Lexical Production. In Bybee, J. and Hopper, P., editors, *Frequency and the emergence of linguistic structure*, pages 229–254. John Benjamins, Amsterdam.

Jurafsky, D., Bell, A., Gregovy, M., and Raymond, W. D. (2001). The effect of language model probability on pronunciation reduction. In *Proc. ICASSP*, pages 801–804.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson, 2nd edition.

Kahn, J. G., Ostendorf, M., and Chelba, C. (2004). Parsing Conversational Speech Using Enhanced Segmentation. In *Proc. HLT/NAACL*, pages 121–128.

Karaiskos, V., King, S., Clark, R. A. J., and Mayo, C. (2008). The Blizzard Challenge 2008. In *Proc. Blizzard Challenge Workshop*.

Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *Proc. MAVEBA*, pages 59–64, Firenze, Italy.

Kawahara, H., Katayose, H., de Cheveigne, A., and Patterson, R. D. (1999a). Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. In *Proc. Eurospeech*, Budapest, Hungary.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigne, A. (1999b). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3-4):187–207.

Kessens, J. M., Wester, M., and Strik, H. (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication*, 29(2):193–207.

Kim, J. (2004). *Automatic Detection of Sentence Boundaries, Disfluencies, and Conversational Fillers in Spontaneous Speech*. Msc thesis, University of Washington, Seattle.

King, S. and Karaiskos, V. (2009). The Blizzard Challenge 2009. In *Proc. Blizzard Challenge Workshop*.

King, S. and Karaiskos, V. (2012). The Blizzard Challenge 2012. In *Proc. Blizzard Challenge Workshop*, Portland, Oregon, USA.

King, S. and Karaiskos, V. (2013). The Blizzard Challenge 2013. In *Proc. Blizzard Challenge Workshop*, Barcelona, Spain.

Klein, D. and Manning, C. (2003). Accurate unlexicalized parsing. In *Proc. ACL*, pages 423–430, Sapporo, Japan.

Kominek, J. and Black, A. W. (2003). CMU ARCTIC databases for speech synthesis. Technical report, Carnegie Mellon University.

Koriyama, T., Nose, T., and Kobayashi, T. (2010). Conversational Spontaneous Speech Synthesis Using Average Voice Model. In *Proc. Interspeech*, number September, pages 853–856, Makuhari, Japan.

Koriyama, T., Nose, T., and Kobayashi, T. (2011). On the Use of Extended Context for HMM-based Spontaneous Conversational Speech Synthesis. In *Proc. Interspeech*, number August, pages 2657–2660, Florence, Italy.

Koriyama, T., Nose, T., and Kobayashi, T. (2012). An F0 Modeling Technique Based on Prosodic Events for Spontaneous Speech Synthesis. In *Proc. ICASSP*, pages 4589–4592, Kyoto, Japan.

Lambert, T., Braunschweiler, N., and Buchholz, S. (2007). How (Not) to Select Your Voice Corpus: Random Selection vs. PhonologicallyBalanced. In *Proc. SSW*, pages 264–269, Bonn, Germany.

Laserna, C. M., Seih, Y.-T., and Pennebaker, J. W. (2014). Um... Who Like Says You Know: Filler Word Use as a Function of Age, Gender, and Personality. *Journal of Language and Social Psychology*, 33(3):328–338.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766.

Lickley, R. (1996). Juncture cues to disfluency. In *Proc. ICSLP*, pages 2478–2481.

Liu, X., Gales, M., and Woodland, P. (2014). Paraphrastic language models. *Computer Speech & Language*, pages 1–19.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., and Harper, M. (2006). Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies. *IEEE Trans. Audio, Speech and Language Processing*, 14(5):1526–1540.

Lu, H. and King, S. (2012). Using Bayesian Networks to find relevant context features for HMM-based speech synthesis. In *Proc. Interspeech*, Portland, Oregon, USA.

Madnani, N. and Dorr, B. J. (2010). Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods. *Computational Linguistics*, 36(3):341–388.

Maekawa, K., Kikuchi, H., Igarashi, Y., and Venditti, J. (2002). X-JToBI: An Extended J-ToBI for Spontaneous Speech. In *Proc. ICSLP*, pages 1545–1548, Denver, Colorado, USA.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marcus-Roberts, H. M. and Roberts, F. S. (1987). Meaningless Statistics. *Journal of Educational Statistics*, 12(4):383–394.

Merritt, T., Raitio, T., and King, S. (2014). Investigating source and filter contributions, and their interaction, to statistical parametric speech synthesis. In *Proc. Interspeech*, Lyon, France.

Mikolov, T., Kombrink, S., Deoras, A., Burget, L., and Černocký, J. (2011). RNNLM - Recurrent Neural Network Language Modeling Toolkit. In *Proc. ASRU Demo Session*, Hawaii, USA.

Miyanaga, K., Masuko, T., and Kobayashi, T. (2004). A Style Control Technique for HMM-Based Speech Synthesis. In *Proc. ICSLP*, Jeju, Korea.

Montero, J. M., Gutierrez-Arriola, J., Colas, J., Enriquez, E., and Pardo, J. M. (1999). Analysis and modelling of emotional speech in Spanish. In *Proc. ICPhS*, pages 957–960, San Francisco, CA, USA.

Morise, M., Yokomori, F., and Ozawa, K. (2016). WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, E99-D(7):1877–1884.

Nakamura, M., Furui, S., and Iwano, K. (2006). Acoustic and Linguistic Characterization of Spontaneous Speech. In *Proc. Speech Recognition and Intrinsic Variation Workshop*, Toulouse, France.

Nakamura, M., Iwano, K., and Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2):171–184.

Nakatani, C. H. and Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *The Journal of the Acoustical Society of America*, 95(3):1603–16.

Neubig, G., Akita, Y., Mori, S., and Kawahara, T. (2012). A monotonic statistical machine translation approach to speaking style transformation. *Computer Speech & Language*, 26(5):349–370.

Obin, N., Lanchantin, P., Avanzi, M., Lacheret-dujour, A., and Rodet, X. (2010). Toward Improved Hmm-Based Speech Synthesis Using High-Level Syntactical Features. In *Proc. Speech Prosody*, pages 3–6, Chicago, Illinois, USA.

Odell, J. (1995). *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, University of Cambridge.

Oertel, C., Gustafson, J., and Black, A. W. (2016). Towards Building an Attentive Artificial Listener: On the Perception of Attentiveness in Feedback Utterances. In *Proc. Interspeech*, pages 2915–2919, San Francisco, CA, USA.

O'Shaughnessy, D. (1992). Recognition of hesitations in spontaneous speech. In *Proc. ICASSP*, San Francisco, CA, USA.

Paul, D. B. and Baker, J. M. (1992). The Design for the Wall Street Journal-based CSR Corpus. In *Proc. DARPA Speech and Language Workshop*, Harriman, NY, USA.

Paulo, S. and Oliveira, L. C. (2005). Generation of Word Alternative Pronunciations Using Weighted Finite State. In *Proc. Interspeech*, pages 1157–1160, Lisbon, Portugal.

Petrov, S. and Klein, D. (2007). Improved Inferencing for Unlexicalized Parsing. In *Proc. HLT-NAACL*, Rochester, New York, USA.

Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release).

Pitt, M. a. and Johnson, K. (2003). Using pronunciation data as a starting point in modeling word recognition. In *Proc. ICPSc*, pages 1–5.

Prahallad, K., Black, A., and Mosur, R. (2006). Sub-Phonetic Modeling For Capturing Pronunciation Variations For Conversational Speech Synthesis. In *Proc. ICASSP*, volume 1, pages I–853–I–856, Toulouse, France.

Raymond, W. D., Dautricourt, R., and Hume, E. (2006). Word-internal /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change*, 18:55–97.

Richmond, K., Clark, R., and Fitt, S. (2010). On Generating Combilex Pronunciations via Morphological Analysis. In *Proc. Interspeech*, number September, pages 1974–1977, Makuhari, Japan.

Richmond, K., Clark, R. a. J., and Fitt, S. (2009). Robust LTS rules with the Combilex speech technology lexicon. In *Proc. Interspeech*, pages 1295–1298, Brighton, UK.

Richmond, K., Strom, V., Clark, R. A. J., Yamagishi, J., and Fitt, S. (2007). Festival multisyn voices for the 2007 blizzard challenge. In *Proc. Blizzard Challenge Workshop*, Bonn, Germany.

Sanford, A. J. S. and Molle, J. (2006). Disfluencies and Selective Attention in Speech Processing. In *Proc. AMNLP*, Turku, Finland.

Scharpff, P. J. and van Heuven, V. J. (1988). Effects of Pause Insertion on the Intelligibility of Low Quality Speech. In *Proc. 7th FASE Symposiom*, pages 261–268, Edinburgh, Scotland, UK.

Schröder, M. (2001). Emotional Speech Synthesis : A Review. In *Proc. Eurospeech*, pages 561–564, Aalborg, Denmark.

Schröder, M. (2009). Expressive Speech Synthesis: Past , Present , and Possible Futures. In Tao, J. and Tan, T., editors, *Affective Information Processing*, chapter 7, pages 111–126. Springer London, London.

Schröder, M. and Trouvain, J. (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, 6:365–377.

Shiwa, T., Kanda, T., Imai, M., Ishiguro, H., and Hagita, N. (2008). How quickly should communication robots respond? In *Proc. HRI*, page 153, New York, USA.

Shiwa, T., Kanda, T., Imai, M., Ishiguro, H., and Hagita, N. (2009). How Quickly Should a Communication Robot Respond? Delaying Strategies and Habituation Effects. *International Journal of Social Robotics*, 1(2):141–155.

Shriberg, E. (1996). Disfluencies in SWITCHBOARD. In *Proc. ICSLP*, pages 11–14, Philadelphia, PA, USA.

Shriberg, E. E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California at Berkeley.

Shriberg, E. E. (1999). Phonetic consequences of speech disfluency. In *Proc. ICPhS*, pages 619–622.

Shriberg, E. E. (2001). To errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). TOBI: A Standard for Labelling English Prosody. In *Proc. ICSLP*, pages 12–16, Banff, Canada.

Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. (2013). Parsing with Compositional Vector Grammars. In *Proc. ACL*, Sofia, Bulgaria.

Stolcke, A., Shriberg, E., Plauche, M., Bates, R., Ostendorf, M., and Lu, Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proc. ICSLP*, pages 2247–2250.

Stolcke, A., Zheng, J., Wang, W., and Abrash, V. (2011). SRILM at Sixteen: Update and Outlook. In *Proc. ASRU*, Hawaii, USA.

Sturt, P., Sanford, A. J., Stewart, A., and Dawydiak, E. (2004). Linguistic focus and good-enough representations: An application of the change-detection paradigm. *Psychonomic Bulletin & Review*, 11(5):882–888.

Sundaram, S. and Narayanan, S. (2002). Spoken language synthesis: Experiments in synthesis of spontaneous dialogues. In *Proc. IEEE Speech Synthesis Workshop*, pages 203–206, Santa Monica, USA.

Sundaram, S. and Narayanan, S. (2003). An empirical text transformation method for spontaneous speech synthesizers. In *Proc. Eurospeech*, pages 1221–1224, Geneva, Switzerland.

Suni, A. and Vainio, M. (2008). Deep syntactic analysis and rule based accentuation in text-to-speech synthesis. In *Proc. TSD*, pages 535–542, Brno, Czech Republic.

Tachibana, M., Yamagishi, J., and Masuko, T. (2005). Speech Synthesis with Various Emotional Expressions and Speaking Styles by Style Interpolation and Morphing. *IEICE Tansactions on Information and Systems*, E88-D(11):2484–2491.

Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, pages 1315–1318, Istanbul, Turkey.

Tokuda, K., Zen, H., and Black, A. W. (2002). An HMM-Based Speech Synthesis System Applied To English. In *Proc. IEEE TTS Workshop*, Santa Monica, USA.

Tomalin, M., Wester, M., Dall, R., Byrne, B., and King, S. (2015). A Lattice-Based Approach to Automatic Filled Pause Insertion. In *Proc. Disfluencies in Spontaneous Speech*, Edinburgh, Scotland, UK.

Trouvain, J., Schmidt, S., Schröder, M., Schmitz, M., and Barry, W. J. (2006). Modelling personality features by changing prosody in synthetic speech. In *Proc. Speech Prosody*, Dresden, Germany.

Trouvain, J. and Schröder, M. (2004). How (Not) to Add Laughter to Synthetic Speech Laughter in Human Interactions. In *Proc. Workshop on Affective Dialogue Systems*, pages 229–232, Kloster Irsee, Germany.

Trouvain, J. and Truong, K. P. (2012). Comparing non-verbal vocalisations in conversational speech corpora. In *Proc. International Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, pages 36–39, Istanbul, Turkey.

van Bael, C. (2007). *Validation, Automatic Generation and Use of Broad Phonetic Transcriptions*. PhD thesis, Radboud University Nijmegen.

van Bael, C., Boves, L., van den Heuvel, H., and Strik, H. (2006). Automatic Phonetic Transcription of Large Speech Corpora. In *Proc. LREC*, pages 4–11, Genoa, Italy.

van Bael, C., van Den Heuvel, H., and Strik, H. (2007). Validation of phonetic transcriptions in the context of automatic speech recognition. *Language Resources and Evaluation*, 41:129–146.

van Son, R. J. J. H. and Pols, L. C. W. (1999). An acoustic description of consonant reduction. *Speech Communication*, 28:125–140.

Villavicencio, F., Bonada, J., Yamagishi, J., and Pucher, M. (2013). Efficient Pitch Estimation on Natural Opera-Singing by a Spectral Correlation based Strategy. Technical report.

Watts, O., Yamagishi, J., and King, S. (2010). The role of higher-level linguistic features in HMM-based speech synthesis. In *Proc. Interspeech*, number September, pages 841–844.

Werner, S., Eichner, M., Wolff, M., and Hoffmann, R. (2004). Toward Spontaneous Speech Synthesis - Utilizing Language Model Information in TTS. *IEEE Transationcs On Speech and Audio Processing*, 12(4):436–445.

Werner, S. and Hoffmann, R. (2006). Pronunciation Variant Selection for Spontaneous Speech Synthesis - A Summary of Experimental Results. In *Proc. Speech Prosody*, Dresden, Germany.

Werner, S. and Hoffmann, R. (2007). Spontaneous Speech Synthesis by Pronunciation Variant Selection - A Comparison to Natural Speech. In *Proc. Interspeech*, pages 1781–1784, Antwerp, Belgium.

Wester, M., Aylett, M., Tomalin, M., and Dall, R. (2015a). Artificial Personality and Disfluency. In *Proc. Interspeech*, Dresden, Germany.

Wester, M., Corley, M., and Dall, R. (2015b). The Temporal Delay Hypothesis: Natural, Vocoded and Synthetic Speech. In *Proc. Disfluencies in Spontaneous Speech*, Edinburgh, Scotland, UK.

Woodsend, K. and Lapata, M. (2011). Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proc. EMNLP*, pages 409–420, Edinburgh, Scotland, UK.

Wu, Z., Valentini-Botinhao, C., Watts, O., and King, S. (2015). Deep neural networks employing Multi-Task Learning and stacked bottleneck features for speech synthesis. In *Proc. ICASSP*, pages 4460–4464, Brisbane, Australia.

Yamagishi, J. (2006). *Average-Voice-Based Speech Synthesis*. PhD thesis, Tokyo Institute of Technology.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Transactions on Audio, Speech and Language Processing*, 17(1):66–83.

Yamagishi, J., Kobayashi, T., Tachibana, M., Ogata, K., and Nakano, Y. (2007). Model Adaptation Approach to Speech Synthesis with Diverse Voices and Styles. In *Proc. ICASSP*, pages 1233–1236, Hawaii, USA.

Yamagishi, J., Onishi, K., and Masuko, T. (2005). Acoustic Modeling of Speaking Styles and Emotional Expressions in HMM-Based Speech Synthesis. *IEICE Transactions on Information and Systems*, E88-D(3):502–509.

Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T. (2003). Modeling of Various Speaking Styles and Emotions for HMM-Based Speech Synthesis. In *Proc. Eurospeech*, pages 2461–2464, Geneva, Switzerland.

Yamagishi, J., Veaux, C., King, S., and Renals, S. (2012). Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology*, 33(1):1–5.

Yamagishi, J. and Watts, O. (2010). The CSTR/EMIME HTS System for Blizzard Challenge 2010. In *Proc. Blizzard Challenge Workshop*.

Yu, Y., Li, D., and Wu, X. (2013a). Prosodic Modeling with Rich Syntactic Context in HMM-Based Mandarin Speeh Synthesis. In *Proc. ChinaSIP*, pages 132–136.

Yu, Y., Zhu, F., Li, X., Liu, Y., Zou, J., Yang, Y., Yang, G., Fan, Z., and Wu, X. (2013b). Overview of SHRC-Ginkgo speech synthesis system for Blizzard Challenge 2013. In *Blizzard Challenge 2013*.

Zechner, K. and Waibel, A. (1998). Using chunk based partial parsing of spontaneous speech in unrestricted domains for reducing word error rate in speech recognition. In *Proc. ACL*, volume 2, pages 1453–1459, Montreal, Quebec, Canada.

Zen, H. (2002). HTS Demo 1.0: Label Overview. http://hts.sp.nitech.ac.jp/archives/1.0/HTS-demo-CMU-Communicator.tar.gz.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., and Tokuda, K. (2007a). The HMM-based Speech Synthesis System Version 2.0. In *Proc. SSW*, pages 294–299, Bonn, Germany.

Zen, H. and Sak, H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *Proc. ICASSP*, pages 4470–4474, Brisbane, Australia.

Zen, H., Senior, A., and Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proc. ICASSP*, pages 7962–7966, Vancouver, Canada.

Zen, H. and Toda, T. (2005). An overview of nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In *Proc. Interspeech*, pages 93–96, Lisbon, Portugal.

Zen, H., Tokuda, K., and Kitamura, T. (2004). An Introduction of Trajectory Model into HMM-Based Speech Synthesis. In *Proc. SSW*, pages 191–196, Pittsburgh, USA.

Zen, H., Tokuda, K., Masuko, T., Kobayasih, T., and Kitamura, T. (2007b). A hidden semi-Markov model-based speech synthesis system. *IEICE Transactions on Information and Systems*, E90-D(5):825–834.

Zhang, Y., Hildebrand, A. S., and Vogel, S. (2006). Distributed Language Modeling for N-best List Re-ranking. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pages 216–223, Sydney, Australia.