THE UNIVERSITY *of* EDINBURGH

Control Strategies For Expressive Text-To-Speech

Atli Thor Sigurgeirsson

Doctor of Philosophy

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

2025

Abstract

When we speak, we convey information beyond the choice of vocabulary—we enhance the message through prosodic augmentation. In addition to its pragmatic functions, such as clarifying intent and guiding the listener's interpretation, prosody is also deployed in expressive speech to communicate emotions, attitudes, and speaking styles. The decision to speak expressively, and how to do so, is shaped by the context in which we speak. Often, multiple renditions may be suitable for a given situation, but choosing the wrong one can be perceived as inappropriate or lead to a misinterpretation of the speaker's intent. Text-To-Speech (TTS) models aim to replicate speech production by mapping written language to speech. Although natural speech can be expressed in various ways, typical TTS models learn a single, likely mapping based solely on the text they are prompted to speak. Through this process, TTS models suppress vocal patterns that deviate from the default speaking manner learned from the training data. Therefore, TTS systems often fail to produce appropriate expressiveness, resulting in overly monotonous speech to the point of impacting overall naturalness.

Instead of relying on text alone to determine a suitable prosodic rendition, controllable TTS models enable users to influence this choice by augmenting various aspects of the speech generation process. In my thesis, I investigate a broad range of strategies for controlling expressive TTS systems: (1) reference-conditioning, in which the generated speech is guided via a speech reference sample; (2) Acoustic Feature Control (AFC), which involves annotators directly manipulating individual features predicted and used by the model; and (3) prompt-based control, where the rendition is described in a natural language text-based instruction. The control strategies I investigate differ regarding the conditioning signal they require and, consequently, how the user interacts with it. The choice of a control method involves trade-offs as the different approaches vary across key aspects of controllability: interpretability of modelled representations, responsiveness to the conditioning signal, specificity of control, and the usability of the chosen method.

Control methods that rely on learning latent representations of prosody, like reference-based models, fail to separate prosody from other factors — such as the identity of the original speaker and linguistic content — from the reference utterance used to create the representation. Resolving this entanglement is challenging due to the opaque nature of the representations. As a result, these methods are often tied to specific input conditions, while sampling new representations from the latent prosody space is

unreliable, leading to reduced naturalness. In contrast, directly manipulating acoustic features provides a more interpretable form of control and can improve the perceived quality of prosody transfer. However, this approach is resource-intensive and technically complex. Experiments also indicate that there is a limit to how complex the task can be if annotators are to improve the predicted rendition. Describing the rendition using natural language offers more accessible control over a limited set of features. Yet, such models typically exhibit uncontrollable variance for the same input text and instruction prompt. I demonstrate how this variability can be reduced by leveraging the model's output distribution through fine-tuning.

Drawing on the differences and limitations of these three control methods, I provide recommendations for the appropriate use of each strategy and suggest ways to address the challenges associated with their implementation.

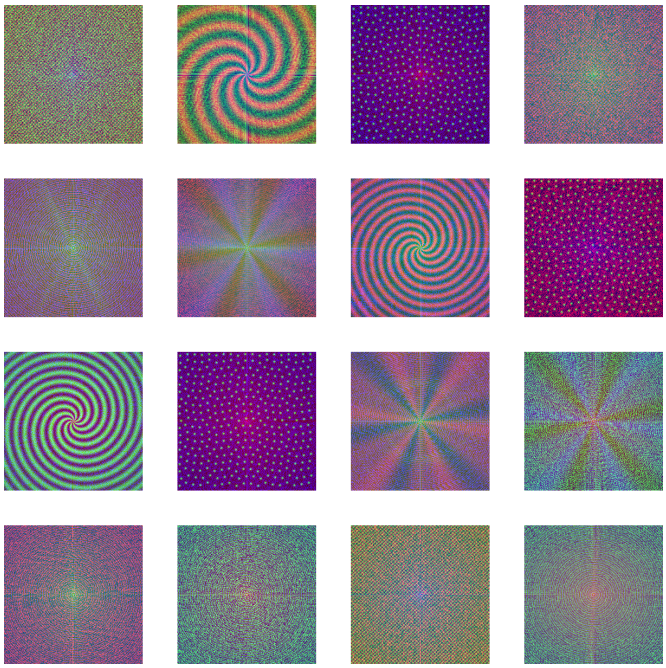
Lay summary

We can say the same thing in many different ways. These differences may lead to people interpreting what we say differently. For example, someone could say "*This game is so much fun!*" in a happy tone, and people would assume that the person genuinely likes the game. If the person said the same exact thing in a sarcastic tone, people would doubt that they enjoy the game. How we say things is based on many different factors: how we are currently feeling, what point we are trying to make, who we are talking to, and what was said before, for example. When we speak, we instinctively choose how to say things based on these and many other factors.

We can also command computers to speak, through a process called Text-To-Speech (TTS). A TTS system takes text as input and produces the corresponding speech. Many modern TTS systems sound astoundingly life-like, so you might imagine that they have learned to produce speech in the same way humans do. But imagine you were asked to perform this task. Given just text, could you always guess an appropriate way of saying it? In many cases, you probably could. But, like in the previous game-based example, there are many cases where you would *have to* base your prediction on some information in addition to just the text.

TTS systems lack the human ability to observe a situation and act appropriately. But, we can help them by turning them into *controllable* TTS systems; systems which can base their prediction on more than just text. In this thesis, I looked into different ways of making TTS systems controllable. There are many different ways to control a TTS, but I considered three: (1) by showing the system how to speak using a human voice sample, (2) by carefully changing values like pitch in specific locations, and (3) describing how it should be said using a text-based description. I did not aim to find "*the best way*" to control a model. Instead, I wanted to address the different weaknesses that these three different methods have. Based on the weaknesses and their characteristics, I suggest when to use which method.

Acknowledgements



I would like to express my gratitude to my supervisor, Simon King, for being an invaluable source of guidance and advice. He has helped me understand that communicating an idea is just as important as the idea itself. I want to thank my many colleagues at the university who have shaped my perspectives while entertaining my silly ideas. In particular, I wish to thank Irene, Eddie, Sarenne, Dan, Johannah and Adaeze, who have made Edinburgh feel like a home away from home over the past five years.

My friends and family back home in Iceland have made this entire journey worthwhile. I am deeply grateful to my parents and my two sisters for their constant support; you mean the world to me. Coming home and finding everything just as it was before moving to Edinburgh makes it all worth it. Lastly, I want to express my heartfelt thanks to my loving partner, Maron. I am immensely thankful to have had someone so supportive, caring, and understanding by my side while I wrote this thesis.

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.

(*Atli Thor Sigurgeirsson*)

Contents

1 Introduction **1**

1.1 Objective and overview 2

1.2 Publications 4

2 Background **5**

2.1 Controllability 5

2.2 Variation in speech 6

2.2.1 Linguistic prosody 7

2.2.2 Portrayal of emotions 11

2.2.3 Speaking style 12

2.2.4 Speaker identity 13

2.3 Text-To-Speech synthesis 13

2.3.1 Historical perspective of speech synthesis control 14

2.3.2 Contemporary speech synthesis 22

2.3.3 Evolution of speech synthesis control 30

I Reference-based Control **33**

3 Introduction and background **35**

3.1 Introduction 35

3.2 Prosody transfer 36

3.2.1 Perspectives on prosody in PT 38

3.2.2 The foundational architecture 39

3.2.3 Evaluation strategies 40

3.2.4 Challenges 43

3.2.5 Extended architectures 44

3.3 Global style tokens . . . 49
3.4 Variational autoencoders . . . 50
3.5 Focus of presented work . . . 51

4 Pilot study: a parallel prosody transfer model 53
4.1 Proposed model and data . . . 54
4.1.1 Reference encoder . . . 54
4.1.2 Speaker encoder . . . 54
4.1.3 Acoustic model and training . . . 55
4.1.4 Data . . . 56
4.2 Model evaluation . . . 57
4.2.1 Conditioning on mel-spectrogram . . . 58
4.2.2 Conditioning on a normalised representation of fundamental frequency . . . 58
4.3 Conclusion . . . 60

5 A training strategy for feature disentanglement 61
5.1 Research objective . . . 61
5.2 Finding prosodically similar pairs . . . 64
5.2.1 Text-based method . . . 64
5.2.2 F_0-based method . . . 64
5.3 Model training . . . 65
5.4 Evaluation . . . 65
5.4.1 Utterance selection . . . 65
5.4.2 Model evaluation . . . 66
5.5 Results . . . 67
5.5.1 RQ 5-1: Can we automatically find prosodically-informative utterances in a speech corpus? . . . 67
5.5.2 RQ 5-2: Does the proposed training regime mitigate the issues of feature entanglement and source-speaker leakage? . . . 69
5.5.3 RQ 5-3: Can a model trained with different yet prosodically similar references retain the level of prosody transfer demonstrated by a baseline prosody transfer model? . . . 73
5.6 Post-hoc reflections . . . 75
5.6.1 Speaker classifier importance . . . 76
5.6.2 Transfer capacity . . . 78

5.6.3 Choice of reference encoder inputs . . . 80
5.7 Conclusion . . . 81
5.7.1 Transferability of representations . . . 82

6 The role of style in speaker identity judgements 85
6.1 Research objective . . . 85
6.2 Stimuli creation . . . 87
6.3 Evaluation . . . 89
6.3.1 Participants . . . 89
6.3.2 Metrics . . . 89
6.4 Results . . . 89
6.4.1 Pre-study . . . 89
6.4.2 RQ 6-1: Can perceived gay voice be successfully modelled by an end-to-end voice-cloning TTS model? . . . 90
6.4.3 RQ 6-2: Which steps in a speech synthesis pipeline have the biggest impact on the fidelity of gay voice? . . . 93
6.5 Conclusion . . . 94

7 Discussion 95

II Acoustic Feature Control 97

8 Introduction and background 99
8.1 Introduction . . . 99
8.2 Background . . . 101
8.2.1 Model refinement through perceptual feedback . . . 102
8.2.2 Acoustic feature control models . . . 104

9 Controllable speaking styles using a large language model 107
9.1 Research objective . . . 107
9.2 Model architecture . . . 110
9.3 Modification method . . . 111
9.4 Prompt construction . . . 113
9.5 Experimental setup . . . 116
9.5.1 The tasks . . . 116
9.5.2 Model training . . . 117

9.5.3 Evaluation of naturalness . . . 117
9.5.4 Evaluation of appropriateness . . . 118
9.6 Results . . . 118
9.6.1 The neutral discourse task (H9-1a and H9-1b) . . . 118
9.6.2 Speaking style generation (H9-1c and H9-1d) . . . 120
9.6.3 Expressive conversational speech (H9-1c and H9-1d) . . . 121
9.7 Conclusion . . . 124

10 Prosody transfer with a human-in-the-loop **127**
10.1 Research objective . . . 127
10.2 Proposed method . . . 130
10.2.1 Baseline model architecture . . . 130
10.2.2 Human-in-the-loop approach . . . 131
10.3 Experimental setup . . . 133
10.3.1 Baseline model training . . . 133
10.3.2 Collection of Human-in-the-loop edited samples . . . 134
10.3.3 Evaluation of edited samples . . . 135
10.4 Results . . . 136
10.4.1 RQ 10-3: Is the suggested control scheme conducive for HitL interaction? . . . 136
10.4.2 RQ 10-1: Can human-in-the-loop participants improve the perceived quality of cross-text prosody transfer? . . . 137
10.4.3 RQ 10-2: Can HitL participants maintain the prosodic similarity with the reference? . . . 140
10.4.4 Post-hoc analysis of user interaction . . . 142
10.5 Conclusion . . . 144

11 Discussion **147**

III Prompt-Based Control **149**

12 Introduction & background **151**
12.1 Introduction . . . 151
12.2 Background . . . 152
12.2.1 Prior influence of LLMs in TTS . . . 152
12.2.2 Types of control . . . 153

12.2.3 Model architecture . . . 154
12.2.4 Evaluation . . . 156
12.2.5 Limitations and challenges . . . 157

13 A strategy for control feature discovery 159
13.1 Research objective . . . 159
13.2 Chapter overview . . . 161
13.3 Baseline models . . . 162
13.3.1 Model architecture and data . . . 162
13.3.2 The two baseline models . . . 163
13.3.3 Baseline model evaluation . . . 164
13.4 Feature discovery and incorporation . . . 169
13.5 Fine-tuning results . . . 172
13.5.1 Fine-tuning of T3 . . . 172
13.5.2 Fine-tuning of T3-emotion . . . 176
13.6 Conclusion . . . 179

14 Discussion 181

15 Conclusions 183
15.1 Method summary . . . 183
15.2 Method comparison . . . 185
15.2.1 Reference-based models . . . 185
15.2.2 Acoustic feature control . . . 187
15.2.3 Prompt-based control . . . 188
15.3 Final remarks . . . 189

A Pilot study: a parallel prosody transfer model 191

B A training strategy for feature disentanglement 193
B.1 Participant instructions for first listening test . . . 193
B.2 Participant instructions for second listening test . . . 193
B.2.1 Prosodic similarity . . . 193
B.2.2 Speaker similarity . . . 194
B.2.3 Naturalness . . . 194

C The role of style in speaker identity judgements 195

C.1 Instructions for listening tests 195
C.1.1 First listening test 195
C.1.2 Second listening test 195

D A human-in-the-loop approach to improving cross-text prosody transfer 197
D.1 Target texts 197
D.2 Reference texts 197

E Controllable speaking styles using a large language model 199
E.1 Instruction prompt 199

F A fine-tuning strategy for discovering controllable features 205
F.1 Instruction prompt 205

Bibliography 207

List of Figures

2.1 Demonstration of the voder at the 1939 world fair in New York 16
2.2 Gunnar Fant demonstrates his OVE I synthesiser 18
2.3 A demonstration of how OVE II reads in parameters 19
2.4 How the TTS task is framed in statistical parametric speech synthesis 23
2.5 Model diagram for an attention-based encoder-decoder TTS model . . 24
2.6 The Tacotron inference process 26
2.7 Architecture diagram for FastSpeech 2 29
2.8 Interpretation of how TTS controllability has evolved over time 30
2.9 A general conditioning method for TTS 31

3.1 The prosody transfer task 37
3.2 The model proposed in Skerry-Ryan et al. (2018) 40
3.3 Daft-Exprt training explanation 47
3.4 Style token selection using reference-conditioning 50

4.1 Model architecture proposed in Chapter 4 55
4.2 F_0 contours generated by the model proposed in Chapter 4. 59

5.1 Prosodic similarity of selected utterances. 68
5.2 QQ-plots confirming approximate normality of residuals 70
5.3 Speaker classification rates reported in Sigurgeirsson and King (2023) 72
5.4 Prosodic-similarity results reported in Sigurgeirsson and King (2023) 74
5.5 Effect of speaker classification sub-loss on mean F_0 77
5.6 Effect of speaker classification sub-loss on F_0 alignment 77
5.7 Effect of embedding capacity on mean F_0 78
5.8 Effect of embedding capacity on F_0 alignment 79
5.9 Effect of reference encoder input on mean F_0 80
5.10 Effect of reference encoder input on F_0 alignment 81

6.1 Flowchart for TTS pipeline used in Sigurgeirsson and Ungless (2024) 88
6.2 Mean *gay voice* ratings for all evaluated speakers. 91
6.3 the correlation between similarity and level of *gay voice* 92
6.4 Mean gay voice rating of both speaker groups in each clip type. 93

9.1 Overview of method proposed in Sigurgeirsson and King (2024) . . . 111
9.2 The role of the LLM in Sigurgeirsson and King (2024) 115
9.3 F_0 variation in the neutral discourse speech task 121
9.4 Preference results for the speaking style generation task 122
9.5 Preference results for the dialogue task. 123
9.6 F_0 variation in the expressive conversational speech task 124

10.1 HitL-based latent exploration for prosody transfer 128
10.2 The HitL-based method proposed in Maurya and Sigurgeirsson (2024) 129
10.3 How HitL-feedback is employed during model inference 131
10.4 A screenshot of the UI employed in Maurya and Sigurgeirsson (2024) 134
10.5 HitL participant effort distributions 136
10.6 Distributions of HitL-participant control-inputs 137
10.7 MOS of original and HitL-adjusted utterances 139
10.8 Prosodic-similarity results for original and HitL-adjusted utterances . 141
10.9 The relationship between HitL effort and the quality of the output . . 142
10.10 The modifications to F_0 and duration suggested by HitL participants . 143
10.11 HitL-participant agreement in F_0, energy, and duration modifications . 144

12.1 Model architecture proposed in Lyth and King (2024) 155

13.1 Overview of method proposed in Sigurgeirsson and King (2025) . . . 161
13.2 Effect of sampling hyperparameters on output distribution 166
13.3 Speaker similarity of synthesised Ingrid to ground truth speakers . . . 168
13.4 Synthesised vs. real similarity of all speakers 169
13.5 PCA visualisation for all Wav2Vec2.0 layers 171
13.6 Clustering based on Wav2Vec2.0 embeddings from synthetic speech . 175
13.7 Overview of 1000 predicted F_0 contours 177
13.8 Large-scale PCA visualisation of Wav2Vec2.0 embeddings 178
13.9 PCA correlation with acoustic features 179

15.1 The controllability profile of reference-based models 186

15.2 The controllability profile of AFC models 187
15.3 The controllability profile of prompt-based models 188

List of Tables

4.1 Training corpus speaker information 57
4.2 Result summary for all models 58

5.1 MOS results for synthetic and ground-truth utterances 69
5.2 Prosodic-similarity results for all models 73
5.3 F_0 contour alignment between synthetic and ground-truth utterances . 75
5.4 Duration differences between output and reference 79

6.1 Naturalness and *gay voice* ratings 90

9.1 Naturalness results for the neutral discourse task 119
9.2 Preference results for both evaluated tasks 121

10.1 Naturalness, speaker similarity, and appropriateness results 138

13.1 Comparison of ground-truth and synthetic speakers 167
13.2 Classification result after first fine-tuning stage 174
13.3 Classification result after second fine-tuning stage 176

A.1 Model hyperparameters for the proposed model 191

Frequently Referenced Architectures

Daft-Exprt *A Prosody Transfer (PT) Text-To-Speech (TTS) architecture proposed in Zaïdi et al. (2022).* 46–49, 53, 60, 63, 65, 66, 69–76, 78–83, 117, 128, 130, 132, 133

FastSpeech 2 *A transformer-based Acoustic Feature Control (AFC) architecture proposed by Ren et al. (2020).* 25, 27–30, 46, 53, 54, 56, 57, 60, 99, 100, 104, 105, 107–110, 116, 117, 122, 124, 147, 184

ParlerTTS *A prompt-based TTS architecture based on Lyth and King (2024).* 153, 156, 161–163, 167, 169

Tacotron *Sequence-to-sequence architecture proposed by Wang et al. (2017).* 24–27, 29, 30, 39, 44, 46, 49, 53, 106

Acronyms

AFC Acoustic Feature Control. iii, xxi, 3, 29, 97, 100, 101, 105–107, 127, 130, 147–149, 151, 152, 181, 184, 187, 188

AL Active Learning. 101, 102

ASR Automatic Speech Recognition. 165

CD Cepstral Distortion. 41

CWE Contextual Word Embedding. 152

DAC Descript Audio Token. 155

DNN Deep Neural Network. 22, 23

DTW Dynamic Time Warping. 40, 42, 64, 65, 75, 77, 78

FFE F_0 Frame Error. 40, 42, 57, 58, 156

FiLM Feature-wise Linear Modulation. 47, 48, 60

G2P Grapheme-To-Phoneme. 20, 23, 24, 26, 27, 57

GPE Gross Pitch Error. 40–42

GRU Gated Recurrent Unit. 54

GSP Gibbs Sampling with People. 103, 104, 128, 129

GST Global Style Token. 36, 49, 50, 96, 103, 109, 117

HCI Human-Computer Interaction. 5

HDI Highest Density Interval. 68, 72, 120–122, 137, 138

HitL Human-in-the-Loop. 100–107, 125, 127–132, 134–145, 147, 187

HMM Hidden Markov Model. 22, 23

IML Interactive Machine Learning. 101, 105

LLM Large Language Model. 100, 107–109, 111–116, 124, 125, 130, 151–157

LME Linear Mixed-Effects. 69–71, 73, 90, 91, 93, 118, 138, 140

LRT Likelihood Ratio Test. 70, 140

MAE Mean Absolute Error. 28, 48

MCD Mel-Cepstral Distortion. 40–42, 67, 156

MCMCP Markov Chain Monte Carlo with People. 103, 104

MFA Montreal Forced Aligner. 28, 56, 65

MFCC Mel-Frequency Cepstral Coefficient. 41

ML Machine Learning. 1, 5, 6, 14, 22, 26, 30, 101, 189

MOS Mean Opinion Score. 43, 66, 67, 69, 89, 118, 119, 135, 136, 138, 139

MSE Mean Square Error. 28, 56

MUSHRA Multi Stimulus with Hidden Reference and Anchor. 42, 43, 67, 69, 73, 74, 135, 136, 140, 141

NFFE speaker-Normalised F_0 Frame Error. 57

NLP Natural Language Processing. 1, 27, 101

NN Neural Network. 1, 2, 22–24

PAT Parametric Artifical Talker. 17

PCA Principal Component Analysis. 161, 162, 169, 170, 172, 173, 177, 179

PESQ Perceptual Evaluation of Speech Quality. 156

PT Prosody Transfer. xxi, 33, 35–40, 42–47, 49, 51, 53, 54, 57–68, 71, 73–76, 78, 81–83, 85, 95, 96, 99–101, 109, 117, 127–131, 133–136, 138, 140–143, 147, 181, 183–186, 188, 197

ReLU Rectified Linear Unit. 28, 191

RNN Recurrent Neural Network. 23, 25–27

RVQ Residual Vector Quantisation. 155, 156

S2S Sequence-To-Sequence. 23–25, 27

SI-SDR Scale-Invariant Signal-to-Distortion Ratio. 156

SLM Speech Language Model. 27, 154, 155, 157, 162, 188

SNR Signal-to-Noise Ratio. 156

SPSS Statistical Parametric Speech Synthesis. 22, 23

SR Speaking Rate. 57

TTS Text-To-Speech. iii, v, xxi, 1–3, 6–8, 13, 18–27, 29–31, 33, 35, 39, 42, 43, 49, 51, 53–56, 62, 69, 73, 87, 88, 90, 93–95, 97, 99–111, 117, 119, 122, 125, 128, 129, 131, 134, 142, 145, 147, 149, 151–154, 156, 157, 159, 160, 162, 181, 183, 184, 187–190

UI User Interface. 134

VAE Variational Autoencoder. 49–51

VDE Voicing Decision Error. 40–42, 156

WER Word Error Rate. 165–167

Chapter 1

Introduction

Have you ever seen someone smash their keyboard? Or someone rapidly click the same button without apparent effect? Maybe you have seen someone desperately whack their computer monitor. All of these people share a common frustration: their digital tool of choice failed to fulfil their intended purpose.

We are surrounded by tools in our daily lives, many of which are incredibly helpful. Increasingly, the tools we rely on are complex, opaque systems based on Neural Networks (NNs). Most people are aware of the immense leap in capabilities that these systems provide when compared to their older counterparts. In particular, generative Machine Learning (ML) models have made waves in recent years. That should come as no surprise, since they seem to effectively tackle any generative modality we task them with. They can generate coherent, long-form text, produce lifelike images, and high-fidelity music and speech, for example. Because of their performance capabilities, generative models are no longer limited to isolated functional use cases. Language models have evolved from their original role in analytical Natural Language Processing (NLP) tasks to become generalized solvers capable of handling a wide range of complex and diverse problems. Image generation is no longer a party trick, as anyone who has studied the Edinburgh Fringe posters in recent years can attest. Text-To-Speech (TTS) models do not "*sound like computers*" anymore, they can maintain a realistic dialogue with their users, employing natural vocal expressions and prosody.

Because generative models have become so powerful, they now rely less on instructions to fulfil the user's intent. In many cases, they *just* predict a sufficiently appropriate output given the limited contextual information they are provided with. But sometimes

they do not. Since NNs don't have a keyboard to smash or a button to rapidly click, what do we do to *control* them when this occurs? That is the focus of this thesis, specifically on the control of TTS models.

But what does it mean to control the process of generating speech? When *we* speak, we do not produce an explicit formulation of how we are going to say something. We *just say it*. Neither do we offer other conversation participants multiple different renditions to pick from; we just generate an appropriate one, based on instinct. From that perspective, we demand more from TTS models than we do from ourselves when we seek to control their outputs. So why should we? Some might also look at the current TTS landscape and deduce that these models have become so good that controlling them is no longer an essential requirement. But those arguments do not hold.

People hold different beliefs about *how things should be said*, but TTS models lack the contextual knowledge required to make *consistently* accurate judgements of how to say things *appropriately*. So, for the foreseeable future, these models will remain dependent on external instruction when they make the wrong prediction. As these models become better, they will presumably make fewer such errors. So over time, this dependency might fade away. But the opposite is true. As errors become less frequent as a result of employing increasingly complex and advanced methods, it becomes more difficult to analyse why they occur: we lose the ability to *interpret* the system. Controllability and interpretability are interlinked concepts: having control over the generation process implies an understanding of how the prediction is made. I argue that effective control is necessary to accurately interpret complex generative models.

Just like the frustrated user pounding a keyboard, we experience friction when generative systems misfire. The sophisticated systems we employ today have abstracted away the levers necessary to guide them. The frustration we experience stems from our innate sense of agency: we desire the ability to affect outcomes when we deem it necessary. This thesis explores control applications for TTS and what it means to create those levers: mechanisms that let us intervene, steer, and ultimately understand these systems when they fall short of our communicative intent.

1.1 Objective and overview

In this thesis, we are concerned with the control of *expressive* TTS models — systems designed to generate speech that varies in aspects such as timbre, emotion, prosody,

and speaking style (Tan et al., 2021). Chapter 2 outlines the motivation for controllability, the specific aspects of speech we aim to provide control over, and how methods for controlling speech synthesisers have evolved over time. The remainder of the thesis comprises three content parts, each describing a different means to control an expressive TTS model:

Part I - **Reference-based control**, which aims to guide aspects of the generated speech using a reference utterance.

Part II - **Acoustic feature control**, where separate acoustic features predicted by the model can be controlled.

Part III - **Prompt-based control**, which conditions speech generation on a natural language description of the desired output.

Each strategy is characterised by its *conditioning signal modality*: reference-based models are controlled using a speech signal, control of Acoustic Feature Control (AFC) models is based on adjustments to acoustic features, and prompt-based models are controlled through text descriptions. Each method comes with its own set of advantages and limitations, which influence the types of tasks for which they are best suited. Much of the research in this thesis seeks to address the known shortcomings associated with each approach:

> **RQ 1-1:** *How can method-specific limitations be addressed?*

This thesis does not aim to suggest a single best control strategy for expressive TTS. The suitability of a control method ultimately comes down to the task in which the method is employed. Instead, the thesis aims to answer:

> **RQ 1-2:** *When should each control method be employed?*

Each content part concludes with a discussion specific to its control strategy. Finally, Chapter 15 provides usage recommendations for each method, based on their controllability characteristics and limitations.

1.2 Publications

This thesis comprises work from five papers published during the course of my PhD:

Chapter 5: *Do Prosody Transfer Models Transfer Prosody?* at the 2023 International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Sigurgeirsson and King, 2023)

Chapter 6: *Just Because We Camp, Doesn't Mean We Should: The Ethics of Modelling Queer Voices* at Interspeech 2024 (Sigurgeirsson and Ungless, 2024)

Chapter 9: *Controllable Speaking Styles Using a Large Language Model* at the 2024 International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Sigurgeirsson and King, 2024)

Chapter 10: *A Human-in-the-Loop Approach to Improving Cross-Text Prosody Transfer* at Interspeech 2024 (Maurya and Sigurgeirsson, 2024)

Chapter 13: *RepeaTTS: towards feature discovery through repeated fine-tuning* at the 13th Speech Synthesis Workshop (Sigurgeirsson and King, 2025)

Chapter 2

Background

2.1 Controllability

In social settings, our sense of agency is determined by our sense of *control* over outcomes (Friston et al., 2013). Based on their current state, a person interacts with the social setting to influence a desired result. The world around us offers varying degrees of controllability — some aspects we can influence, while others remain beyond our control, despite our desire to change them. This notion of control, central to our sense of agency, has a formal counterpart in technical disciplines.

In the context of Machine Learning (ML), the understanding of system controllability is rooted in control theory. The controllability of a dynamic system is defined as the system's ability to achieve a desired configuration over a finite time horizon, given appropriate control signals (Kalman, 1960). Therefore, for generative ML models, controllability can be viewed as the model's ability to base its generation on a human-supplied control signal. The controllability of a ML system must be evaluated based on the task that the system aims to solve. From the control theory perspective, a method that requires three interactions to yield an acceptable solution and another that requires 300 interactions are both considered controllable. From a Human-Computer Interaction (HCI) perspective, the former method is considered more controllable as it is more *usable* (Gould and Lewis, 1985).

The controllability of a system also has implications in terms of the system's interpretability. The interpretability of a system describes the degree to which a human can understand the internal mechanics of the system (Gilpin et al., 2018), such that they

are able to understand the cause for a decision made by the system (Miller, 2019). Interpretable ML models have many benefits compared to *black-box* counterparts. Interpretable models can be more efficiently analysed when the enumeration of all possible outputs is not tractable, and they provide a better framework for identifying ethical concerns regarding fairness, privacy, and objective misalignments (Doshi-Velez and Kim, 2017, for example). For discriminative ML models, the interpretability of the model is typically linked to how easy it is to understand how the model made a particular classification. From this perspective, decision trees (Morgan and Sonquist, 1963) and Bayesian rule lists (Letham et al., 2015), for example, are often considered interpretable because the model classification can be understood in terms of a sequence of transparent steps. In the context of generative ML models, interpretability is often linked to the model's ability to independently control *disentangled* factors (Ross et al., 2021).

Today, interactions with ML systems, which offer varying degrees of controllability, influence the social context in various ways. From a philosophical perspective, providing control over ML systems may contribute to people's sense of agency in general. From a more practical standpoint, controllability can make generative ML models more usable. We should assume that any generative ML model is capable of producing an undesirable output, and methods need to be put in place to regain control and to steer the model generation in those cases (Kieseberg et al., 2023). This need is particularly pertinent for Text-To-Speech (TTS) models (Section 2.3), where the same input text may correspond with a wide range of plausible speech outputs (Section 2.2), while a listener might consider only a subset of them appropriate for a given context.

2.2 Variation in speech

Speech is inherently variable. Social and idiosyncratic biological differences yield perceptual differences between different speakers (Bradlow et al., 1999). The way we speak depends on who we are talking to, the environment and context, what we are discussing, and our emotional state. Within-person variability spans many dimensions, such as prosody which plays a crucial role in communication (Lavan et al., 2019). The deployment of prosody, whether intentional or not, enables speakers to convey emotion (Cole, 2015, p. 13), social context, and intent beyond the literal meaning of the words spoken (Wilson and Wharton, 2006). For some contexts, listeners expect

a particular manner of speech (*speaking style*) which may not be suitable for other contexts (Eskenazi, 1993). Therefore, any given sequence of words can correspond to a broad spectrum of different but perfectly reasonable (i.e. expected, given the communication context) vocal renditions. However, the word sequence itself only provides a limited indication of what constitutes a reasonable rendition. TTS models, which aim to convert text into spoken language, typically rely solely on textual input to generate speech. As a result, they lack direct access to many factors that drive vocal variation.

Gaining control over variation — due to differences between speakers and within-speaker variation shaped by conversational context — may thus provide a way to guide expressive output in dimensions the model cannot infer from text alone. The current section provides a discussion of four sources of variation in speech that are later investigated throughout the thesis: (1) linguistic prosody, (2) emotional state, (3) speaking style, and (4) speaker identity. Of course, there are relationships between these features; prosody shapes speaker identity judgements (Helander and Nurminen, 2007) and different speaking styles correspond with differences in prosody (Abe, 1997), for example. But, in the context of TTS, the objective is often to achieve independent control of one or more of these sources of variation.

2.2.1 Linguistic prosody

A speaker can convey the same message in several ways (Wilson and Wharton, 2006). Prosody describes how an utterance is said to encode information in the speech signal, in addition to the lexical content. But, prosody should not be viewed as something added on top of a message comprising basic segments, such as syllables. Instead, prosody forms an integral part of speech production and often provides a meaningful contribution to the message itself (Clark et al., 1995, p. 328).A speaker may employ prosody to encode a particular interpretation, aid comprehension, or convey attitude and emotion (Cole, 2015, p. 13) for example. Some prosodic renditions are appropriate in a given context while others are not; choosing the right one can therefore be critical to communicate a particular meaning (Wagner and Watson, 2010).

Prosody serves two broad communicative functions, which can be grouped according to their underlying objectives: (1) the *affective function*, whose goal is to communicate emotions, attitude, and state-of-mind; and (2) the *augmentative function*, which aims to aid comprehension, influence a particular understanding (Taylor, 2009, p. 13-19), and

to signal structure and information status (Cole, 2015). These are just a few examples of the roles that these two functions serve, whose objectives are different and realised in different ways. Affective prosody is realised over long spans in a given communication context, influencing overall perception. Augmentative prosody is, on the other hand, highly coupled to the choice of lexical items (Wagner and Watson, 2010).

In this section, we focus on augmentative prosody, which is often just referred to as prosody or *linguistic prosody*. Prosody is not realised in the same manner across languages. The discussion that follows aims to provide an overview of how important aspects of prosody are used in the English language. There are many different theories of prosody, but this section focuses on just three fundamental phenomena of linguistic prosody: stress, phrasing, and intonation. As we will see, different uses of these three prosodic phenomena can yield different interpretations. Ultimately, prosody functions within a communicative context shared between speaker and listener, relying on mutual assumptions about how and when to use it effectively (Clark et al., 1995, p. 330); whereas a TTS system has limited access to contextual cues which inform the choices made for these prosodic phenomena. Therefore, providing control over synthesised stress, phrasing, and intonation enables manual adjustments that can alter the meaning of the message.

2.2.1.1 Acoustic correlates of prosody

The principal acoustic correlates of prosody are duration, fundamental frequency (F_0), and energy, which serve both segmental and suprasegmental functions (Clark et al., 1995, p. 322). However, these acoustic features are employed for many other, non-prosodic purposes. At the segmental level, these cues help encode phonological information, such as distinguishing long and short vowels in English. Over longer suprasegmental spans, they assume additional roles, contributing to prosodic functions such as intonation, stress, and rhythm. Many linguistic contrasts are not due to phonemic differences, but rather to these suprasegmental features, which operate across multiple segments (Roach, 1989, p. 36). Because prosody often reflects higher levels of linguistic organisation, it can be difficult to disentangle from other long-term characteristics (Clark et al., 1995, p. 329-331). Moreover, listeners interpret these cues contextually, making judgments about the speaker's personality, gender, or age, particularly based on pitch (Clark et al., 1995, p. 323).

2.2.1.2 Stress

Stress refers to the relative emphasis put on syllables or words within speech, typically marked by a relative increase in loudness, pitch, and duration. Often, all of these acoustic cues are used together to signal stress, but not necessarily (Roach, 1989, p. 73). Stress contributes to word recognition by helping listeners segment words from continuous speech (Cole, 2015, p. 4) and by providing contrast between homophones (Arvaniti, 2020, p. 3). For example, the noun form of *subject* is associated with stress on the first syllable while the verb form is associated with stress on the latter.

Prominence, or the use of stress to mark a particular word in a phrase as more salient than others, contributes to overall interpretation through a disambiguation function (Taylor, 2009, p. 137). For example, any word in the phrase:

> *he walked across the field*

can be made prominent to deliver a different interpretation; e.g. "***he*** *walked across the field*" to indicate that it was ***he*** who walked as opposed to *someone else*, or "*he* ***walked*** *across the field*" to emphasise that the person ***walked*** rather than, say, *ran*. Out of pitch, duration and loudness, pitch is regarded as the most salient and loudness the least salient, or at least the most inconsistent, acoustic correlate of prominence (Clark et al., 1995, p. 332-335).

2.2.1.3 Phrasing

Phrasing, or *grouping*, refers to how speech is organised into units, comprising multiple words, to clarify sentence structure (Frazier et al., 2006). Different prosodic phrases are separated by a prosodic boundary, which is realised through pauses, pitch resets, and final lengthening (Cole, 2015). Prosodic phrasing aids listeners in parsing and, consequently, understanding the meaning of the message. In some cases, prosodic boundaries can be predicted by the syntactic structure of the text. For example, in:

> *From a technical point of view* | *the plan seemed sound*

the indicated prosodic phrase boundary would be considered natural, although other boundaries could be valid as well. But in many cases, prosodic phrasing is more *flat* than syntactic phrasing, which is inherently hierarchical (Taylor, 2009, p. 113). Consider:

> *Where is the can* | *which was next to the bottle* | *that was part of the gift* | *that I received on my birthday*

Here, the prosodic phrasing overrides syntactic phrasing to yield a rhythm more typical for English speech than would result from phrase boundaries based on syntactic structure. Phrasing can also be employed for disambiguation (Frazier et al., 2006). For example, in the phrase "*Steve or Sam and Bob will come*", the placement of a prosodic boundary can yield different interpretations:

Steve | *or Sam and Bob will come*

would indicate that either Steve is coming, or both Sam and Bob are coming. A different placement can be employed:

Steve or Sam | *and Bob will come*

In this case, either Steve or Sam are coming while Bob definetely is.

2.2.1.4 Intonation

Intonation is probably the most complex component of English prosody, and sometimes the terms *intonation* and *prosody* are used interchangeably. Intonation is primarily marked by changes in pitch during the course of a speech act (Taylor, 2009, p. 121) and principally realised through three functional aspects of pitch: (1) *tone*, e.g. a rising/falling tone over a period of time; (2) the *tone placement*, determining which segment is intoned; and (3) the *tone group structure*, or how a tone develops over consecutive syllables, and the combination of more than one tone within an utterance (Clark et al., 1995, p. 358).

Intonation can be used to make syllables or words relatively more prominent than others (the *accentual function* of intonation). As with many other aspects of prosody, there isn't a clean-cut boundary between realisations of stress and intonation since tones are typically realised on locations of lexical stress (Clark et al., 1995, p. 360). Intonation is used to mark boundaries to inform of grammatical structure and sentence type (the *grammatical function*). Intonation is, for example, used to distinguish between statements and questions. A rising tone on the last lexical stress typically indicates a question, while a falling one indicates a statement:

She lent him her (↗ car) vs. *She lent him her (↘ car)*

The former rendition would indicate a question, while the latter indicates a statement. Intonation, therefore, serves a vital role in discourse, indicating new or given information, to cue for a response and what sort of a response is expected (the *discourse*

function) (Roach, 1989, p. 136-137). Beyond these augmentative functions, intonation also plays an important role in conveying emotions and attitudinal factors (the *attitudinal function*) (Roach, 1989, p. 138-142).

2.2.2 Portrayal of emotions

Emotions are conveyed through several communication *channels* but most importantly: body language, facial expression, and vocal expression (Wirth and Schramm, 2005). Numerous experiments have shown that emotions can be communicated through affective prosodic channels alone (e.g., Davitz and Davitz, 1959; Burns and Beier, 1973; Costanzo et al., 1969). Expressions of emotions and attitudes are complex and can be performed both intentionally and involuntarily. Emotions do not necessarily reflect an interaction with an interlocutor; they can also be directed towards the topic of the conversation, a third party, or something else external to the communication context. All of this shapes how affective prosody is performed and perceived (Roach, 1989, p. 137-138). Like linguistic prosody, different emotional renditions may yield different interpretations of the same linguistic message. Consider, for example:

> *I would really like to see you right now.*

A joyful emotional rendition might indicate excitement for a reunion, while a rendition characterised by an angry tone could indicate a desire for a confrontation.

Vocal correlates of emotional expressions have been investigated in a large number of studies. Portrayals of emotion influence long-term features such as mean pitch, pitch range, and formant frequencies (e.g., Frick, 1985; Juslin and Laukka, 2003). In addition to spectral features, changes in duration, voice quality, and intensity are fundamental to the perception of emotions and attitudes (Clark et al., 1995, p. 331). Often, vocal expressions of emotions are analysed in terms of *fundamental emotions*, such as anger, arousal, fear, happiness, sadness, and boredom (Scherer et al., 2003, Chapter 23). Acoustic differences between polar opposite emotions (e.g. happy vs. sad) are particularly salient. Expressions of happiness are linked to a general rise in pitch and a faster rate of articulation (Frick, 1985; Juslin and Laukka, 2003), although this is not truly universal (Tartter and Braun, 1994; Drahota et al., 2008). On the other end, sad speech exhibits a low average pitch, a monotonous intonation, and a slow speaking rate (Murray and Arnott, 1993). Perception of emotions is also shaped by the emotional *intensity* in which the utterance is rendered, and the conversational context in which the emotion is portrayed (Bachorowski and Owren, 1995).

2.2.3 Speaking style

Speakers may choose to employ an appropriate *speaking style* for a given communication context, yielding perceptually salient differences that don't necessarily reflect the speaker's emotion or identity. Traditionally, the term has been used to distinguish only between two different classes of speaking styles: *read/laboratory speech* and *spontaneous speech* (Hirschberg, 2000). But there are many other different contexts where listeners expect a particular style of speech. For example, a TV news anchor may use a particular speaking style to give a neutral, authoritative impression. At the same time, an advertisement reader may choose to sound energetic to appeal to the listener. Speaking styles may be determined by the corresponding text content in some cases. However, several external factors influence the choice of speaking style, including the number of listeners (*conversational speech* vs. *public speaking*), the age of the listener (e.g. *child-directed speech*), and the external environment in which communication is performed (*casual* vs. *formal*), for example (Eskenazi, 1993). However, a speaking style can also reflect an unconscious habit, potentially influenced by the speaker's anatomy, or a purposeful strategy for conveying a specific personality (Clark et al., 1995, p. 329). Although speaking styles differ across languages, most languages possess means of conveying a range of speaking styles (Clark et al., 1995, p. 330)

The perception of speaking style is shaped by, at the very least, the conversational context (Eskenazi, 1993), prosodic and spectral characteristics (Abe, 1997) and both verbal and non-verbal content (Gustafson et al., 2021). Attempts have been made to develop a taxonomic view of speaking styles, classifying different styles based on *style axes*: (e.g. formal vs. non-formal / clear vs. unclear / read vs. spontaneous) and task-specific roles (e.g. interviewing, news reading, sports announcer) (Eskenazi, 1993; Prateek et al., 2019), and certain speaking styles can be distinguished by differences in prosody. For example, formal public speaking tends to be very rhythmic, while spontaneous speech is less so (Roach, 1989, p. 102-104). It is generally accepted that speaking style describes information-rich patterns that influence a speaker's choice of the *global* deployment of prosody (Wang et al., 2018). These patterns determine the suprasegmental behaviour of the acoustic correlates of prosody; voice quality, speaking rate, F_0 variation, and loudness (Eskenazi, 1993; Yamagishi et al., 2005; Hazan and Baker, 2010), for example. Listeners can perceive distinctions arising from different speaking styles and employing a speaking style in the wrong context may be perceived as unnatural or inappropriate (Van Santen et al., 2013, Chapter 33).

2.2.4 Speaker identity

Speech is inherently influenced by the speaker, with anatomical differences between individuals resulting in perceptible acoustic variation. The frequency at which the vocal folds vibrate is dictated by their length, mass, and tension (Schweinberger et al., 2014). These anatomical differences of the vocal folds also affect voice quality and, therefore, *timbre*. Similarly, differences in the vocal tract shape and dimensions affect the speaker's formant structure, although learned behaviours may also contribute to these inter-speaker differences (Schweinberger et al., 2014). Difference in mean F_0 is often thought to be the most discriminative feature influencing speaker similarity judgements (Baumann and Belin, 2010). However, listeners also rely on a wide range of acoustic parameters relevant to voice quality when making similarity judgements (Kreiman et al., 1992).

The perception of speaker identity is not only shaped by individual acoustic features. According to Bricker and Pruzansky (1966), listeners identify voices by constructing a map of the speaker's phonemic inventory, and this process becomes more effective as linguistic variability increases. There is also a strong correspondence between how speakers use prosody and the perception of their identity (Helander and Nurminen, 2007). Socio-demographic variables also influence speaker identity judgements: regional accents and perceived social origin (Rakić et al., 2011), perceived age (Mulac and Giles, 1996), and other perceived personality traits, such as trustworthiness (Vukovic et al., 2011). Yet, research has faced challenges in defining a consistent set of acoustic cues that reliably signal identity across different voices (Kreiman and Sidtis, 2011); in other words, the acoustic features that distinguish voices within one speaker group may not generalise well, as they are often derived from limited, person-specific data. While many researchers emphasise between-person variability as a discriminative feature for distinguishing different speaker identities, an important facet of the perception of speaker identity is the within-person variability of speech production (Lavan et al., 2019).

2.3 Text-To-Speech synthesis

TTS is the automatic conversion of written text to spoken language. Its central goal is to produce a natural and comprehensible spoken rendition of the text. The development of speech synthesis spans from its pre-electronic beginnings in the 18th century

to the present day, marked by ML-driven solutions. Here, I provide a concise background history of speech synthesis development (Section 2.3.1) and the contemporary perspective (Section 2.3.2) to contextualise the work conducted in this thesis.

This account is primarily characterised by the various mechanical, electrical, and digital methods devised to generate an acoustic analogue of human speech. But, as this brief overview demonstrates, the history of speech synthesis is also shaped by a fundamental question: *how do you control a speech synthesiser?*

2.3.1 Historical perspective of speech synthesis control

For most of human history, speech was an act that only humans were capable of performing. But, stories of legendary *automata* capable of speech date back to at least the 8th century BC. Perhaps the oldest preserved mention of a machine capable of speech comes from Homer's Iliad:

> *He [Hephaestus] spoke, and from the anvil rose, a huge, panting bulk, halting the while, but beneath him his slender legs moved nimbly . . . but there moved swiftly to support their lord handmaidens wrought of gold in the semblance of living maids. In them is understanding in their hearts, and in them speech and strength, and they know cunning handiwork by gift of the immortal gods.*
>
> (Book XVIII, p. 410-420 Homer and Murray, 1924)

Hephaestus, the Greek god of blacksmiths and technology, is depicted as constructing mythical automata capable of movement, strength, and the autonomous production of sounds to facilitate communication. As automata, they self-governed their production of speech, requiring no external guidance. These references introduce a vision of technological intentions while, in this case, the realisation is ascribed to the Gods (Paipetis, 2008, p. 79). It wasn't until two millennia later that human efforts to construct such a machine were realised, with the help of another man named Homer.

2.3.1.1 Pre-electronic era

Spanning as far back as the 15th century, various real automata had been invented to create the illusion of mechanised speech (Hoffmann, 2019). These devices, or *talking heads*, did not generate speech on their own; rather, a concealed operator would speak into a tube to animate the device. *Speaking dolls*, or ventriloquist figures as they are more commonly known, were also used to the same illusionary effect. These fraudulent speech synthesis attempts captivated observers — occasionally to the detriment of

the inventors, some of whom were accused of witchcraft and faced legal proceedings:

> *The doll, that the inventor had in his arms, was able to reply to all questions asked. This fact came to the knowledge of [sic] Inquisition ... During his trial, that was turning for the worse, the accused asked the judges to question directly the doll. Doctors cross-examined on religious problems and it answered every question in a very satisfactory way. Inquisitors were so pleased that released to it a Catholicism certificate [sic]. Eventually it turned out that the inventor was a ventriloquist who had cheated everybody.*
>
> (Giannini, 1999, p. 2535)

Although these precursor devices were not true speech synthesisers, they revealed a long-standing interest in acquiring such a device.

The first genuine speech-synthesis device is often credited (Hoffmann, 2019) to Wolfgang von Kempelen's late-18th-century apparatus (Kratzenstein, 1781). The Austro-Hungarian Von Kempelen, who is perhaps more famous for his invention of the *mechanical Turk* (Clark et al., 1999, p. 126-165), believed that mechanical speech production should replicate the human production of speech. Therefore, a speaking machine should comprise mechanics matching the human anatomy; a mechanical lung, glottis, and a mouth would be required (Von Kempelen, 1791, p. 398). Von Kempelen's device consisted of kitchen bellows for lungs, a bagpipe reed for the glottis, and a bell from a clarinet was attached for a *"mouth"*. The machine provided limited control: compressing the bellows supplied airflow through the machine, and the rigid bell allowed for rudimentary control of vowel production (Hoffmann, 2019).

2.3.1.2 Electronic era

It wasn't until the 20th century, during the *electronic era* of speech synthesis, that machines were capable of generating intelligible speech. Early efforts focused on the electronic production of formants. One such device, developed by John Q. Stewart, comprised two resonant circuits, each tuned to a formant, alongside a telephone receiver to convert the current into sound waves, and an *interrupter* to modulate the current (Stewart, 1922). The machine was operated by adjusting the frequency of the interrupter, enabling changes to the frequencies of the first two formants. Understandably, the machine was limited to vowel production.

It was Homer Dudley's invention of the *vocoder* (Dudley, 1939) that marked a turning point in this era, and of speech synthesis as a whole. Speech, like any other sound, is a

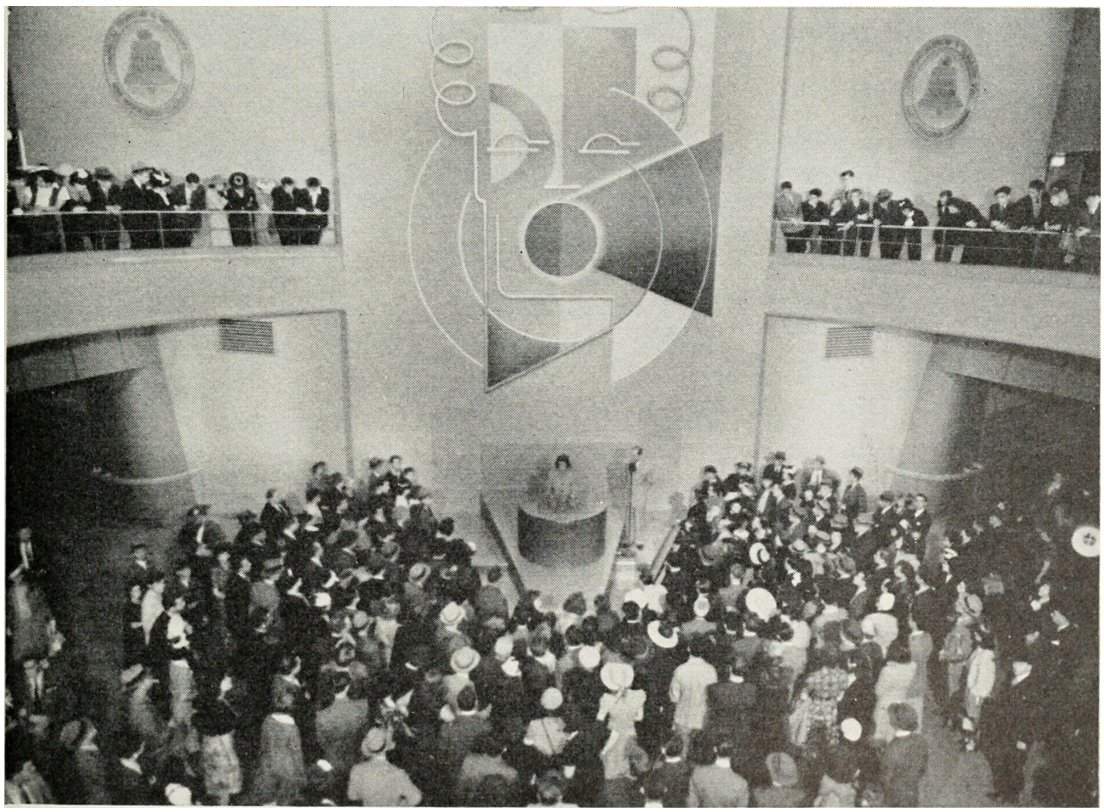


Figure 2.1: A trained operator demonstrates the *voder* at the 1939 world fair in New York (Williams, 1940)

continuous sound wave propagated through a medium. But Dudley's vocoder enabled the approximation of speech in terms of slower-varying individual parameters instead. The vocoder was an electronic device that could generate a parametric description of an incoming speech signal (encoding) and then reverse this process, converting a parametric description back into an approximation of the speech waveform (decoding). Given an input speech signal, the device produced approximations of the fundamental frequency (0-250 Hz) and the composition of energy across 10 spectral bands, ranging from 250 Hz to 2950 Hz. These approximations would then drive an electronic synthesiser to produce audible speech (Dudley, 1939).

But Dudley's vocoder could not synthesise arbitrary speech; it could only resynthesise the speech input to the system. How such a system should be controlled to synthesise a desired output was the obvious next question. Dudley's answer was his next invention, the *voder* (Dudley et al., 1939), which was based on the parametric principles of his vocoder. The voder was a human-operated machine that replaced the speech input of the vocoder with control inputs supplied by a human operator. A pedal controlled the fundamental frequency, and a *piano-like* keyboard modulated the gain of a bandpass

filter to control the 10 spectral bands. Using a *wrist bar*, the voice source could be changed for voiced (oscillating wave) or unvoiced (random noise) sounds (Klatt, 1987, p. 741). Operating the voder was non-trivial, and operators required months of training to control it (Dudley et al., 1939). Although voder's intelligibility was *marginal* (Klatt, 1987, p. 741), it was a significant invention that continues to shape the field of speech synthesis to this day.

> "*... he saw impressively the effect it had upon them. That they were hearing something of startling scientific import and profound human interest was obvious from the expressions on their faces, for uniformly they listened in rapt and appreciative attention.*"
>
> – A member of Bell staff describes observers' reactions to the voder (Williams, 1940, p. 63)

The introduction of the *source-filter model* (Fant, 1961) constituted the next major advance in the historical account of speech synthesis. This theory posits that speech can be viewed as the result of an excitation source being linearly filtered, such that some frequency components are increased in amplitude and others are decreased. In human speech production, there is a correspondence, on one hand, between the source and phonation, and between the filter and articulation on the other (Fant, 1961, p. 17). As such, the source corresponds to the vibrations of the vocal folds or turbulent airflow, while the filter corresponds to the vocal tract. The first formant synthesisers that supported dynamic formant control were built on this new theory of speech production (Klatt, 1987, p. 742). One such example is Gunnar Fant's OVE I (Stevens et al., 1953). OVE I generated the first two formants using mechanical resonators. The fundamental frequency, voice source, and the two formants could be controlled using a mechanical arm. The arm could be pulled and pushed over a plane to plot out the two formant frequencies on individual axes. The fundamental frequency and amplitude of the source were controlled with a potentiometer (a *knob*) (Klatt, 1987, p. 742).

Similar to the voder, OVE I required *active* control, where speech is controlled as it is generated. The next version, OVE II (Fant, 1952), allowed for *programmed* control instead, where control of formants was specified before synthesis. OVE II employed a reading system, initially proposed by Walter Lawrence (Lawrence, 1953) (inventor of Parametric Artifical Talker (PAT)), which read in formant trajectories drawn in conductive ink. In this way, the formant frequencies over time could be *programmed*, thus removing the need for active control. OVE II also included mechanisms for synthesising fricatives and nasal consonants were devised (Klatt, 1987, p. 742).

Figure 2.2: Gunnar Fant demonstrates his OVE I synthesiser. A mechanical arm was employed to actively plot out the first two formant frequencies. Image is taken from Watanabe et al. (1979), original photographer unknown.

2.3.1.3 Digital era

Fant had thus demonstrated how speech synthesis could be pre-configured with his OVE II synthesiser. But in his method, each different utterance required its own *program*, written in conductive ink. A long-standing goal was to achieve automatic speech synthesis for any input text, without manual intervention. The dawn of the digital age, which laid the foundation for modern TTS, marks the initial success of systems that achieved this goal. Earlier mechanical or electronic attempts to control synthesisers were phased out as process computers became available (Hoffmann, 2019, p. 17). Early in the digital age, automatic, or *rule-based*, synthesis was developed in three main paradigms, each of which aimed to generate arbitrary utterances from synthesised phonetic segments: (1) rule-based formant synthesis; (2) rule-based articulation synthesis, which aims to produce speech by simulating human articulation; and (3) concatenative speech synthesis (Klatt, 1987, p. 752). The core aim of each paradigm was the same: generate an intermediate *linguistic specification* from text inputs using a rule-based system, which is sufficient to drive a speech synthesiser. Where they

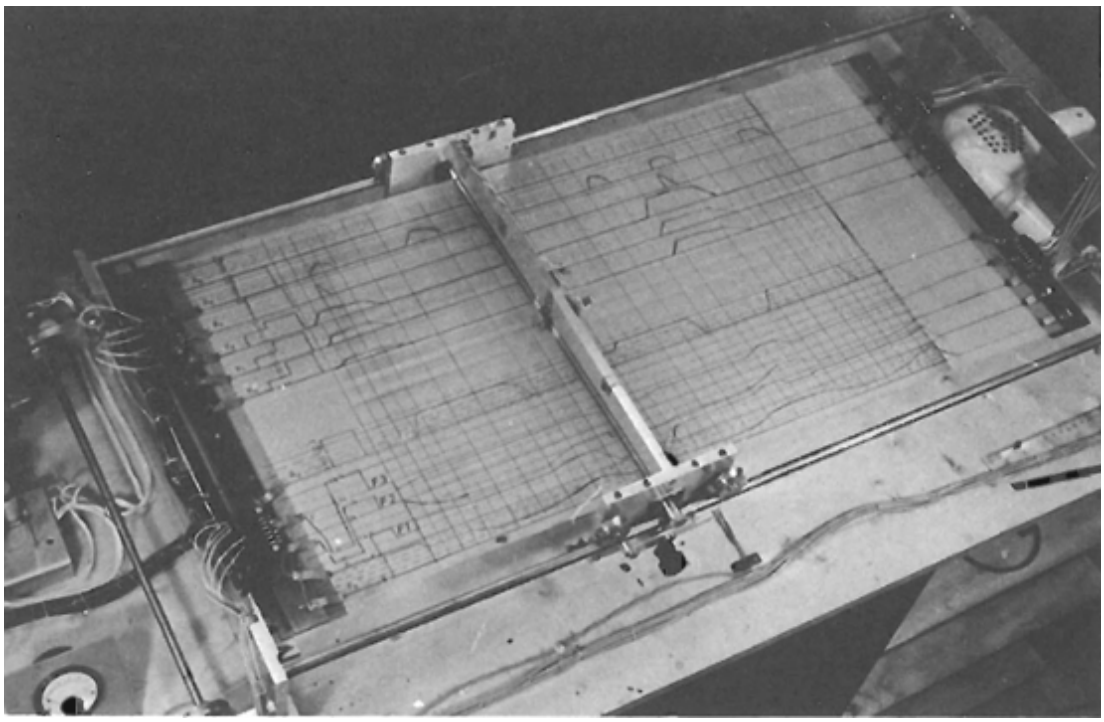


Figure 2.3: The OVE II synthesiser would read in parameter trajectories, including formant frequencies, written in conductive ink. Image is taken from Watanabe et al. (1979), original photographer unknown.

differed was in how speech is synthesised from the linguistic specification. Formant and articulation synthesis had already been extensively studied. Early rule-based approaches enabled automatic synthesis by writing standard rules for, for example, phone formant frequencies and how they should transition between phones (Holmes et al., 1964). It was during this time that the first actual TTS system was developed (Teranishi and Umeda, 1968), based on a rule-based articulation system (Klatt, 1987, p. 49).

The third paradigm, concatenative speech synthesis, is the one least similar to approaches that characterised the electronic age. Enabled by process computers, concatenative speech synthesis creates the desired speech by arranging pre-recorded speech units into the corresponding sequence. *Diphone synthesis*, a type of concatenative synthesis, is achieved through the selection of a diphone sequence from a pre-recorded dataset, and the joining of the sequence to produce speech (Dixon and Maxey, 1968). In its simplest form, a diphone synthesiser simply concatenates the diphones from the inventory. However, such systems are inherently constrained by their fixed diphone inventories: the acoustic properties of each diphone are predetermined, preventing manipulation of fundamental frequency, duration, and other prosodic features. One could expand the inventory to include several variations of each diphone, enabling some

acoustic variation. However, this rapidly inflates the required diphone inventory size (Klatt, 1987, p. 758). This speech synthesis paradigm later evolved into *unit selection* (Taylor, 2009, p. 484-528), which enabled the dynamic selection of speech units, allowing whole words or phrases to be selected if they were present in the inventory, thereby benefiting both intelligibility and naturalness.

These early, simple rule-based systems showed promise, but the limited rules that they employed restricted the naturalness and intelligibility of the output. More would be required than stringing together a sequence of phones to realise the synthesis of perceptually natural speech. Over time, rule-based systems grew more elaborate: new pronunciation and prosody rules were introduced, existing ones refined, and exceptions explicitly defined. The objective was to derive a sufficiently detailed linguistic specification from text input to drive accurate, natural-sounding speech generation. Many types of rules were developed; notably, rules for Grapheme-To-Phoneme (G2P) conversion (text normalisation, stress prediction, and allophone selection), as well as rules for prosody (For example, how F_0, intensity, and duration develop across the utterance) (Klatt, 1987, p. 759-767).

Together, this collection of rules — which dictate how various speech features take their values — is referred to as the *front-end*. Because the front-end features are explicitly modelled, the front-end provides multi-dimensional control. Previously, the entire speech waveform was generated from the same set of limited mechanical or electric inputs. Now, different things could be controlled separately and simultaneously. The method of achieving front-end feature control differed between systems, but typically through a computer console. Some would allow for the adjustment of the phonemic transcription generated by the front-end. Other systems could take hand-drawn fundamental frequency contours as input, enabling control of F_0 across the entire utterance (Klatt, 1987, p. 763).

Developing a front-end capable of generating a *complete* and accurate linguistic specification for any novel text proved to be difficult—such a development required sophisticated syntactic and semantic analysis of the text. Improvements could be made by adding more rules, but these additions would only yield minor, incremental improvements (Klatt, 1987, p. 767). In his TTS systems review published in 1986, Dennis H. Klatt states that no system at the time was capable of automatically performing the semantic analysis required to correctly predict prosodic aspects — such as stress and pause durations — for arbitrary text inputs.

The next approximately twenty years were characterised mainly by advancements in concatenative synthesis, in particular unit selection (Taylor, 2009, p. 484-528). Like the prior rule-based systems, unit selection systems employ a front-end to generate a linguistic specification of the input text. The unit selection system selects the optimal sequence of units based on a *target cost* and *join cost*. Before synthesis, the inventory is labelled with contextual features, for example: phoneme features (e.g. position in syllable), word features (e.g. part-of-speech), and phrase features (e.g. position in major phrase) (Tokuda et al., 2002; Clark et al., 2004). When evaluating the target cost of a unit for selection, the system compares the unit's features with those predicted by the front-end. The join cost reflects how well the chosen units fit together consecutively, e.g. in terms of spectra, F_0, and energy discontinuities across the join (Clark et al., 2004). Methods based on unit selection could generate more natural speech, in general, than their formant and articulation counterparts, and therefore became the focus of speech synthesis research during this period.

It should, however, be noted that early formant synthesisers were able to produce remarkably natural speech. But to achieve this, careful manual adjustment of formant frequencies and amplitudes was required. There were hopes that these manual adjustments could be described in a rule-based system, but the problem of developing a *theory of control* remained (Klatt, 1987, p. 744).

Human parity in the 70s

Several contemporary TTS models are described to have achieved *human parity* in terms of perceived naturalness (e.g., Tan et al., 2024). Although an impressive achievement, some would argue that this milestone had already been achieved by John Holmes in 1972 with his formant synthesiser:

> *...the average listener could not tell the difference between a synthetic and natural sentence when presented with both in sequence.*
>
> Dennis H. Klatt describes the quality of Holmes' formant synthesiser (Klatt, 1987, p. 743)

Many other remarkable synthesis inventions and technologies were developed during the 20th century that have not been mentioned in this concise overview, for example: Klattalk, linear prediction analysis and *Speak-n'-Spell*, and voice conversion using the DECtalk system (Klatt, 1987). However, we now shift the focus to the era of speech synthesis research in which we currently find ourselves.

2.3.2 Contemporary speech synthesis

2.3.2.1 Statistical parametric speech synthesis

The start of the last decade marked the beginning of the current era of speech synthesis, characterised by ML-based methods for synthesis. This development started with the Statistical Parametric Speech Synthesis (SPSS) models, which demonstrated competitive performance compared to systems based on unit selection (King, 2010). early SPSS models were based on Hidden Markov Models (HMMs), where a separate HMM is trained for each *contextualised* phoneme (Taylor, 2009, p. 446-482). Each HMM is trained to learn the statistics of vocoder parameters, corresponding to its contextualised phoneme. Given an input text, the front-end generates contextualised phonemes based on the linguistic specification. The corresponding sequence of trained HMMs is then combined, and a generation algorithm is used to predict the most probable vocoder parameters given the state sequence.

Soon after, in 2013, the first SPSS employing Deep Neural Networks (DNNs) (Bengio et al., 2009) was proposed by Zen et al. (2013). Similar to its HMM-based counterpart, this model predicted vocoder parameters from an input sequence encoded with contextual features. But in Zen et al. (2013) these predictions were performed by a DNN instead of HMMs. The DNN predicted frame-level features from frame-level inputs. During training, the input sequence was upsampled according to an extracted alignment. During inference, durations were explicitly predicted. This DNN-based approach outperformed a similarly sized (in terms of parameters) HMM-based model, marking the start of the currently active field of *neural* TTS. Today's era is characterised by the application of Neural Network (NN) architectures at every stage of the Text-To-Speech pipeline; bringing both benefits and challenges in terms of control.

2.3.2.2 Sequence-to-sequence models

At this point, there was a well-established three-step paradigm to SPSS (Figure 2.4): (1) a front-end analyses the input text to generate a linguistic specification; (2) given the linguistic specification, an acoustic model predicts vocoder parameters; and (3) a vocoder generates a speech waveform from the parameters. But, the naturalness of SPSS was primarily limited by the acoustic modelling and vocoding they employed, compared to unit selection (Black et al., 2007). Vocoders at the time produced *buzzy* speech, attributed to a simplistic, periodic view of the excitation source (King, 2010,

p. 13). Second, HMM-based acoustic models were prone to making inaccurate, over-smoothed vocoder parameter predictions, resulting in muffled speech (Black et al., 2007). Unit selection does not require a vocoder; therefore, it bypasses these issues.

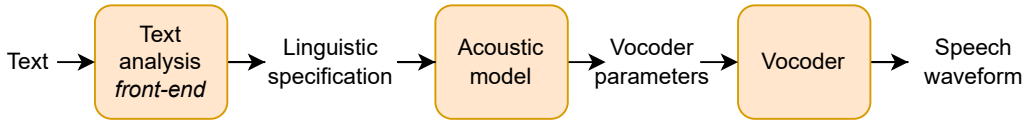


Figure 2.4: In SPSS, TTS is typically viewed as a three step paradigm. A front-end generates a linguistic specification to aid acoustic modelling. An acoustic model generates vocoder parameters from the linguistic specification. Given the vocoder parameters, a vocoder generates a speech waveform.

To some degree, these two issues in SPSS have already been addressed without the use of neural networks. Mixing aperiodic excitation with the periodic excitation source reduces the perceived *buzziness* of the speech waveform (e.g., Kawahara, 2006; Morise et al., 2016). To address over-smoothing, Toda et al. (2005) proposed *global variance*, which aims to adjust predicted vocoder parameters such that they maintain variation observed in real speech. Nonetheless, speech synthesis research focus shifted to NN-based approaches. Various methods have been developed to replace the different steps in the SPSS paradigm with NN architectures instead of, e.g., rule-based front-ends and HMMs-based acoustic models. The previously mentioned first DNN-based approach to TTS (Zen et al., 2013) replaced the acoustic model. Wavenet (van den Oord et al., 2016) revolutionised TTS with a neural autoregressive vocoder architecture, predicting one waveform sample at a time. Different tasks of the front-end were tackled separately using different NNs: prosodic-event detection (Jeon and Liu, 2009), text normalisation (Sproat and Jaitly, 2016), and G2P conversion (Yao and Zweig, 2015), for example.

Ultimately, though, it became clear that employing NNs throughout the TTS in a *unified* pipeline was an effective approach that comes with several advantages. To achieve this, TTS is framed as a Sequence-To-Sequence (S2S) task (Sutskever et al., 2014), where a variable-length input sequence is encoded and then decoded to produce a variable-length output sequence in the target domain. The encoder and decoder are based on Recurrent Neural Networks (RNNs) (Medsker et al., 2001) to handle the variable-length inputs and outputs. But for speech, the difference in the input/output modalities results in a significant discrepancy between the two sequences; only a few words can correspond to >100 mel-spectrogram frames. This difference is a problem for RNNs, which suffer from the *vanishing gradient problem* where training becomes

ineffective for very long sequences (Hochreiter, 1998). So, S2S TTS models typically employ an *attentive* (Bahdanau et al., 2014) recurrent decoder to address this problem (e.g., Wang et al., 2016, 2017; Sotelo et al., 2017; Shen et al., 2018). In short, attention provides a general framework for assigning dynamic weights to elements in an input sequence of a NN. Thus, in S2S TTS, attention allows the decoder to *attend* to each element in the encoded input sequence when making a mel-spectrogram frame prediction. In this way, attention provides the means for learning the alignment between text and mel-spectrogram.

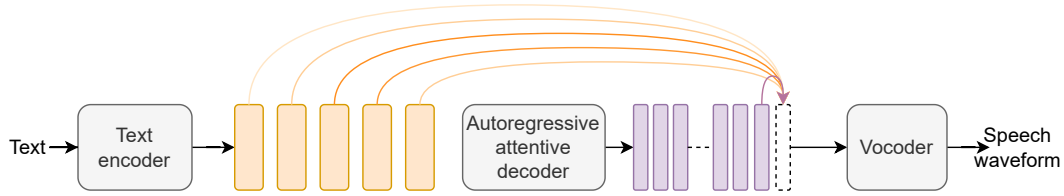


Figure 2.5: A typical attention-based S2S model for TTS comprises a recurrent text encoder and an attentive autoregressive mel-spectrogram decoder. In **Tacotron** (Wang et al., 2017), the prediction of every mel-spectrogram frame is conditional on the previously predicted one, as well as an attention-weighted representation of the encoded input sequence.

Since S2S TTS models employ a unified architecture, they can be trained *jointly* to minimise a single loss term. The unified architecture also allows them to generate speech in an *end-to-end* manner: waveform prediction is made directly from text inputs (e.g., Shen et al., 2018). Reducing the TTS task into this general format comes with several incentives. Most importantly, front-end requirements are reduced, making it easy to extend end-to-end TTS systems to different speakers and languages without the need for a sophisticated linguistic specification. Understandably, learning an end-to-end mapping from text to speech is difficult. Consider, for example, the benefits of a TTS front-end. Typical speech corpora used for TTS training comprise far fewer distinct pronunciations than a pronunciation dictionary. So, training on grapheme inputs requires the TTS model to learn the grapheme-to-phoneme conversion implicitly from a limited number of examples. Compared to a lexicon-based approach, end-to-end TTS models produce less accurate phonetic predictions (Taylor and Richmond, 2019). Therefore, end-to-end models are often trained on phonemised inputs — generated by an auxiliary G2P model — instead of text (e.g., Ren et al., 2019, 2020; Zaïdi et al., 2022). Various S2S TTS architectures have demonstrated the ability to generate high-fidelity natural speech (e.g., Sotelo et al., 2017; Wang et al., 2017; Shen et al., 2018).

Important model architectures

Several TTS model architectures are frequently discussed throughout this thesis. Notably, **Tacotron** (Wang et al., 2017) and **FastSpeech 2** (Ren et al., 2020) were widely adopted and have been influential in the TTS research community. These two architectures, which informed much of my early research and are foundational for Part I, are described here and in Section 2.3.2.3. Other foundational architectures, relevant to the various control methods explored later in the thesis, are discussed in the relevant chapters.

Tacotron

Tacotron (Wang et al., 2017) is a S2S TTS model that internally learns alignment using attention (Bahdanau et al., 2014). **Tacotron** comprises an input sequence encoder (employing text inputs instead of phonemes) and a speech representation decoder. **Tacotron** predicts a linear-scale spectrogram instead of mel-spectrogram. First, text is projected into a sequence of high-dimensional embeddings, using an embedding table. These embeddings are transformed through a series of non-linear layers before encoding, performed with a bidirectional RNN (Rumelhart et al., 1985). The transformation produces a sequence of text-resolution latent embeddings. The spectrogram decoder, based on the same architectural principles as the text encoder, expects inputs at frame resolution. To align these two sequences, **Tacotron** uses an attention-based RNN. At each decoding step, this attention-RNN takes a transformed representation of the previously predicted frame as input. The goal of this RNN is to predict an attention *query* over encoder steps, given the last hidden state of the attention-RNN and its current input. From this query, attention over encoder steps is performed to produce a *context vector*: an attention-weighted sum of the encoder steps.

The context vector is concatenated with the attention-RNN output. Given this combined representation as input, and its previous hidden state, the RNN-decoder predicts spectrogram frames. For **Tacotron**, spectrogram prediction is an auto-regressive process and, therefore, a time-consuming one. Exploiting the short-term stationary nature of speech, it instead predicts two frames ($r = 2$) at a time, thus speeding up inference. In Wang et al. (2017), the model stops generation after a maximum number of steps[1]. In a follow-up paper, a special `<stop>` token is employed to indicate that generation should stop (Shen et al., 2018). **Tacotron** is trained to minimise the $l1$ loss of spectrogram prediction. Wang et al. (2017) demonstrated that the model learns to align text

[1] Presumably, since this is not mentioned in the original **Tacotron** paper.

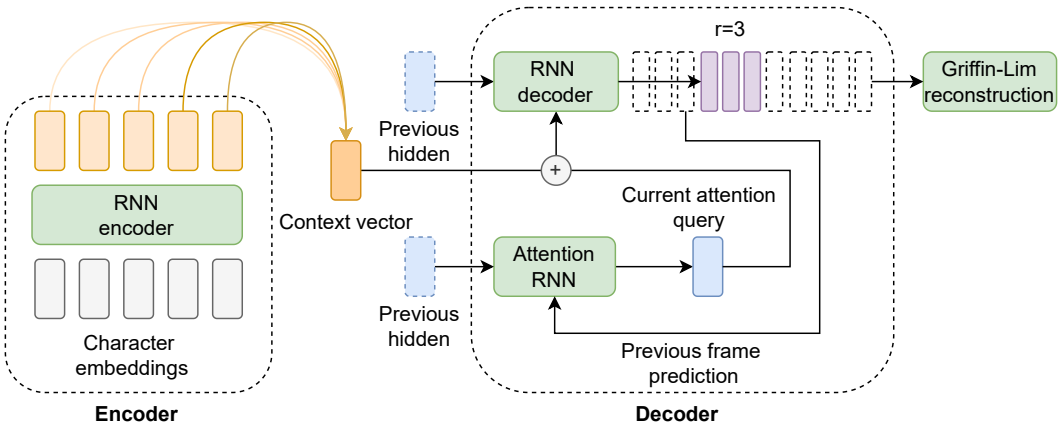


Figure 2.6: A high-level overview of the **Tacotron** inference process. At each decoding step, the **attention RNN** produces an attention query based on the previously predicted frame. This query is used to generate a context vector, an attention-weighted sum of encoder steps. The attention query and context vector serve as inputs to the spectrogram **RNN decoder**, which can be trained to predict $r > 1$ frames at a time.

inputs with spectrogram outputs, achieving very high naturalness for the time. Wang et al. (2017) describe **Tacotron** as an end-to-end architecture when the model is combined with a Griffin-Lim (Griffin and Lim, 1984) reconstruction module. However, this is somewhat misleading, as the model still predicts spectrogram frames.

The **Tacotron** architecture has several notable flaws. As frame prediction is conditional on the prediction made at the previous time step, **Tacotron** is sensitive to even a single false prediction of the `<stop>` token, which can lead to early stopping where speech is cut off prematurely (Shen et al., 2020). **Tacotron** also places no constraints on the alignment, so **Tacotron** exhibits errors stemming from misalignment: long pauses, repetitions, or babbling (Shen et al., 2020). As **Tacotron** is auto-regressive, decoding of speech is performed over hundreds of steps, making both training and inference slow processes. Nonetheless, **Tacotron** became a very popular framework for TTS research (e.g., Skerry-Ryan et al., 2018; Wang et al., 2018; Zhang et al., 2019), as it does not require external tools such as a G2P or an aligner.

2.3.2.3 Transformer-based architectures

Currently, the TTS landscape is very much shaped by the use of *Transformers* (Vaswani et al., 2017). Transformers are a general purpose ML construct based on *self-attention*, which models internal dependencies within an input sequence. Transformers have shaped numerous fields of research and continue to do so. In particular applications

in Natural Language Processing (NLP), for example: text summarisation (Liu and Lapata, 2019), language understanding and question-answering (Devlin et al., 2019), and most famously language modelling (e.g., Devlin et al., 2019; Brown et al., 2020; Ouyang et al., 2022). Whereas models like **Tacotron** employ *cross-attention* between encoder and decoder steps to compute the alignment, self-attention enables modelling long-range contextual dependencies *within* the input sequence itself.

Early transformer-based TTS models typically followed the encoder-decoder paradigm: a transformer-based architecture encodes the input text/phoneme sequence and another transformer-based architecture decodes a mel-spectrogram from the encoded sequence (e.g., Li et al., 2019; Lańcucki, 2021; Ren et al., 2019, 2020). Compared to S2S models based on RNN architectures, transformer-based models can perform *parallel* generation, where the full mel-spectrogram is predicted in a single step (e.g., Lańcucki, 2021; Ren et al., 2019, 2020; Zaïdi et al., 2022). Lately, though, several so-called Speech Language Models (SLMs) have been proposed for TTS. These models frame the speech generation task similarly to language modelling where a decoder-only transformer architecture autoregressively predicts audio *codes* (e.g., Borsos et al., 2023a,b; Wang et al., 2023). These models will be revisited in Part III. But here, **FastSpeech 2** (Ren et al., 2020) which is an illustrative example of the transformer-based encoder-decoder paradigm, is discussed in detail.

FastSpeech-2

Due to its inefficiencies in training and inference, as well as the attention alignment issues outlined in Section 2.3.2.2, many researchers had sought a **Tacotron** replacement architecture which could address both of these problems. FastSpeech provided such a solution (Ren et al., 2019). However, I focus here on the slightly extended version, **FastSpeech 2** (Ren et al., 2019), which serves as a baseline model architecture in many of the studies that follow in this thesis.

FastSpeech 2 is a non-autoregressive transformer-based TTS architecture. It takes phoneme inputs and predicts the mel-spectrogram in parallel. Like typical neural TTS models, **FastSpeech 2** comprises an encoder and a mel-spectrogram decoder. The phoneme encoder consists of four feed-forward *transformer blocks*, a stack of multi-head, scaled dot-product self-attention (Vaswani et al., 2017) and a one-dimensional convolution (LeCun et al., 2015) layer. First, the text is phonemised, using a pre-trained G2P model, and embeddings are generated for each phoneme, using an embedding table. The phoneme encoder generates a sequential, contextualised representation of

the phoneme embeddings, which serves as the input for the so-called *variance adaptor*. The role of this module is to separately predict salient acoustic variance, which is then employed to help the model reconstruct the mel-spectrogram.

In **FastSpeech 2**, the variance adaptor comprises three *variance predictors*, one for each of F_0, energy, and duration. All three predictors share a similar structure, consisting of a series of 1-dimensional convolutional layers with Rectified Linear Unit (ReLU) activations. Each convolutional layer is followed by layer normalisation (Ba et al., 2016) and dropout (Hinton et al., 2012). Although they employ similar modelling strategies, the duration predictor serves the additional role of upsampling the phoneme sequence. Variance prediction is performed in a series, with duration prediction being the first step. The duration predictor takes the contextual phoneme sequence as input and predicts the duration (number of mel-spectrogram frames) of each phone. The prior FastSpeech architecture (Ren et al., 2019) employed a *length regulator*, which upsamples the latent phoneme sequence based on the predicted durations. This length regulator is also employed in **FastSpeech 2**. It is because of the length regulator that **FastSpeech 2** does not require attention to produce alignments between phonemes and acoustic features. After upsampling, the pitch and energy predictors make frame-level predictions for F_0 and energy. Predictions are made within a quantised range for these two features, determined by the minimum and maximum values in the training corpus. The predicted values are projected into high-dimensional embeddings, which are then summed to the upsampled phoneme sequence. It should be noted that in this design, energy predictions are dependent on F_0 predictions since these are modelled in series.

Finally, the mel-spectrogram decoder—consisting of 4 transformer blocks—predicts a mel-spectrogram from the latent variance adaptor output sequence. A linear layer is added on top of the decoder to match the desired number of mel-spectrogram frequency bins. Because the intermediate latent sequence representation has already been upsampled to frame resolution, the mel-spectrogram can be predicted non-autoregressively. During training, the model employs ground-truth values to train the variance predictors. For this, alignments are extracted using Montreal Forced Aligner (MFA) (McAuliffe et al., 2017), l^2-norm of mel-spectrogram frames for energy, and F_0 estimated using the WORLD library (Morise et al., 2016). Each variance predictor is trained to minimise the Mean Square Error (MSE) loss of their respective feature. The model is jointly trained to minimise the variance predictor loss terms as well as the Mean Absolute Error (MAE) of mel-spectrograms.

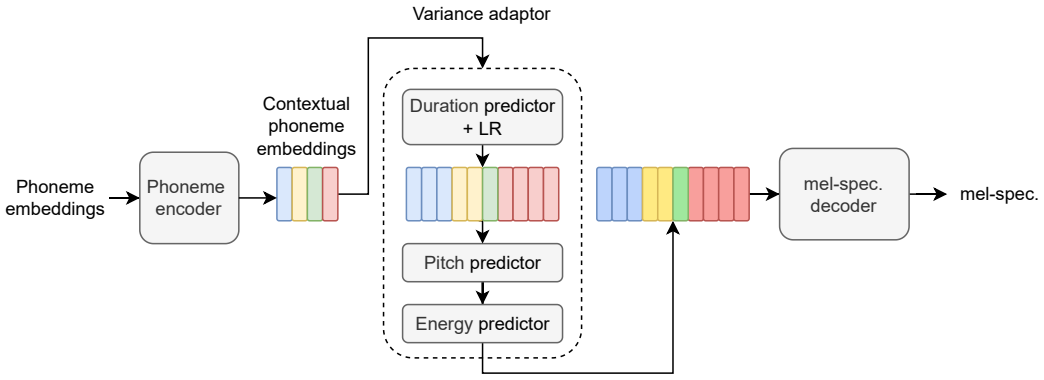


Figure 2.7: Model diagram for **FastSpeech 2** (Ren et al., 2020). The input text is first phonemically transcribed and projected into embeddings. From there, a Transformer-based phoneme encoder generates a sequence of *contextual* phoneme embeddings. The duration predictor and Length Regulator (LR) upsample each embedding to a mel-spectrogram frame resolution. The F_0 and energy predictions are projected and summed to the upsampled sequence. The output of the variance adaptor is then decoded into a mel-spectrogram prediction for the input text.

FastSpeech 2 produces speech that is comparable to **Tacotron** in perceived naturalness (Ren et al., 2020). Since the mel-spectrogram is predicted non-autoregressively, **FastSpeech 2** requires far less time for training and offers more efficient inference. Because phone durations are learned from ground-truth alignments, attention is not required to map phonemes to acoustics, thus resolving the alignment issue. Any input text can yield multiple settings of the three acoustic features learned by **FastSpeech 2**. Explicitly learning how text corresponds to variation in these features is found to be helpful for model training (Ren et al., 2019). But there is an additional benefit of modelling speech in this way. Duration, F_0, and energy are salient acoustic correlates of prosody and are important in the production of intonation, for example. So, modelling these features explicitly may have additional benefits for prosody generation. In a sense, **FastSpeech 2** is also a more *interpretable* architecture than other typical end-to-end TTS architectures: the model prediction is factorised into interpretable acoustic features. Because of how **FastSpeech 2** models speech, it also enables rudimentary control of the acoustic features it models. I will refer to models that enable such control as Acoustic Feature Control (AFC) models, which will be further discussed in Part II.

2.3.3 Evolution of speech synthesis control

Clearly, the field of TTS has been shaped by the incremental but rapid improvements in ML research, and this development has, overall, been for the better. Neural TTS can generate more natural and intelligible speech than any other prior technology (Tan et al., 2021). Neural TTS models are less dependent on linguistic and phonetic expertise when compared to the models of the past, so TTS is available in more languages than ever before. This dependency reduction stems from the capacity of neural TTS systems to integrate multiple stages of the traditional TTS pipeline — such as the front-end, the acoustic model, and in some cases the vocoder — into a single model that leverages learned, contextually rich intermediate representations. But, to some extent, the various benefits of neural TTS have resulted in loss of control over the speech generation process.

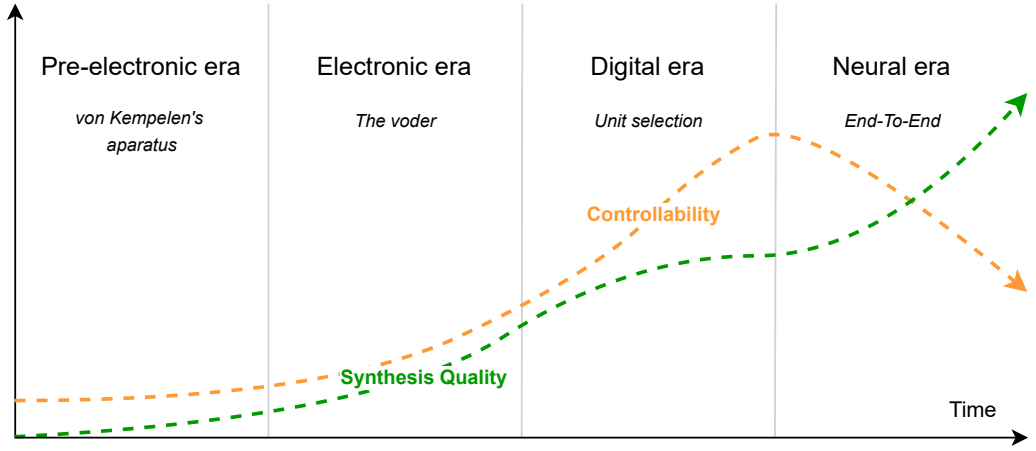


Figure 2.8: An interpretation of how speech synthesis controllability and the quality of synthesis have evolved over time; technology representative of each period is included in italics, time is not to scale.

Consider, for example, the unit selection models of the past. By performing speech generation based on a sophisticated linguistic specification, precise adjustments can be made to the various features included in the specification. Most end-to-end TTS models offer no such control mechanism; the model predicts the most probable speech corresponding to the input text. This lack of control has implications for, for example, prosody generation. Prosody synthesised by most end-to-end TTS models is an *average prosody*, reflecting the broad statistics of the training data (Hodari et al., 2019). End-to-end TTS models do not generally select amongst a wide range of possible prosodic renditions. Models like **Tacotron** and **FastSpeech 2** produce a single most

probable prosodic rendition for each input text. The average prosody these models predict may be suitable in many cases. But they lack the required context knowledge to make judgements about appropriate prosodic variation when it is expected. Employing average prosody when it is not anticipated impairs comprehension (Govender and King, 2018) and perceived naturalness.

However, neural TTS models are flexible architectures, allowing for *conditioning* speech generation on more than just the input text. That is, control of neural TTS models is typically achieved by training the TTS model to base speech prediction on auxiliary, separately encoded features in addition to text input. During inference, the model can be controlled by supplying it with a control input that describes some target model behaviour. Various features of speech may be of interest for control, and this general conditioning strategy has been employed for speaker identity (e.g., Wan et al., 2018; Arik et al., 2017; Jia et al., 2018), prosody (Henter et al., 2018b; Cai et al., 2020; Kenter et al., 2019), speaking style (e.g., Wang et al., 2018; Valle et al., 2020a), and the speaker's accent (e.g., Henter et al., 2018a; Liu et al., 2020) amongst other things.

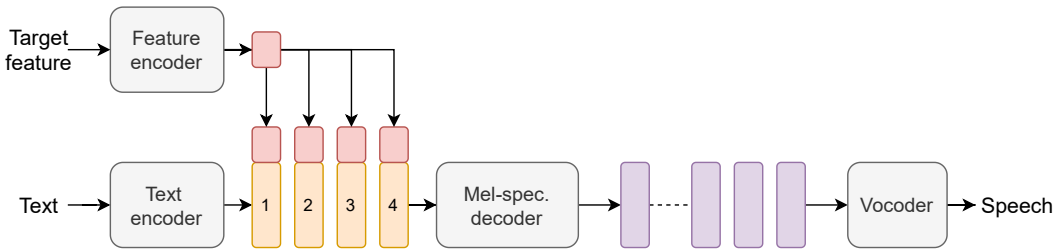
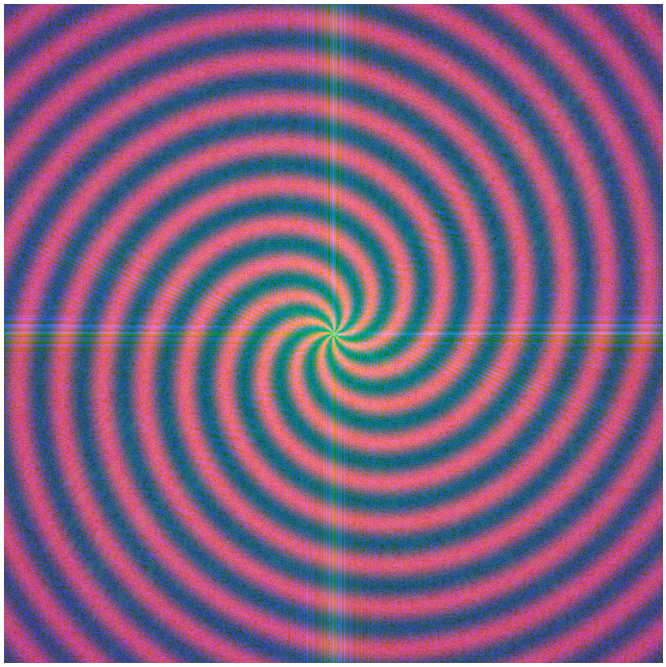


Figure 2.9: Neural TTS architectures can control speech generation via a conditioning signal corresponding to a target feature, e.g. desired speaker/accent/emotion. In this simple illustration, a single embedding (red) is appended across all encoding time steps. However, the conditioning process performed differs between models in terms of where the conditioning signal is introduced within the model and how the signal is combined with other model features.

There are several ways to achieve model conditioning and differences exist in terms of how the target features are encoded, how they are combined with other information signals in the acoustic model, and in which model location conditioning is performed. These differences may reflect what is being controlled and how temporally-precise the control is intended to be. This thesis investigates three distinct control strategies, each defined by the modality of its conditioning signal. Each approach offers unique advantages and limitations, making it more or less suitable for different use cases. The thesis begins with an exploration of reference-based models.

Part I

Reference-based Control



Reference-based control for TTS aims to guide certain aspects of the output speech using speech itself to guide the generation. This specific area of research represents a significant focus of my PhD thesis, in particular, the so-called Prosody Transfer (PT) models. The following part brings together three works conducted over the whole span of my PhD, which is reflected in its length.

Published papers presented in this part:

> **Chapter 5**: *Do Prosody Transfer Models Transfer Prosody?* (Sigurgeirsson and King, 2023)
>
> **Chapter 6**: *Just Because We Camp, Doesn't Mean We Should: The Ethics of Modelling Queer Voices* (Sigurgeirsson and Ungless, 2024)

Chapter 3

Introduction and background

3.1 Introduction

A potentially intuitive approach to controlling the generation of speech is through the speech medium itself. Such control can be achieved using *reference-conditioning* for Text-To-Speech (TTS) where the model is guided via a reference utterance supplied to the model.

Reference-conditioning is a generic control method that is applied across various TTS tasks. In dialogue-based synthesis, the provided reference corresponds to a prior utterance in a conversation (Oplustil-Gallegos and King, 2020). In emotive speech modelling, the reference utterance dictates the emotional tone of the output (van Rijn et al., 2021). Reference-conditioning is also deployed for controlling speaking styles (Wang et al., 2018; Valle et al., 2020a), speaker identities (Eren and Team, 2021) and accents (Henter et al., 2018a). Another key application of reference-conditioning is Prosody Transfer (PT) (Skerry-Ryan et al., 2018), which aims to transfer the prosody from a reference utterance onto generated speech. Like with other applications of reference-conditioning, this is achieved by conditioning speech generation on a learned representation of the reference. PT is sometimes described as a *"say it like this"* task (Skerry-Ryan et al., 2018, p. 1), as it aims to synthesise a target text in the voice of a target speaker while matching the prosodic characteristics of the reference. Much of my work focuses specifically on this particular use of reference-conditioning.

A fundamental requirement for effective reference-based control is the model's ability to separate the control feature from others in the reference signal. In PT, failing to do so

compromises the preservation of the target speaker's identity and the naturalness of the output. When this separation fails, the reference features are said to be *entangled*. The feature entanglement problem, and its underlying causes, are discussed further in Section 3.2.4. Here, I refer to representations that enable such feature separation as *transferable*. Feature entanglement remains a key challenge in reference-based control, and much work has focused on addressing it. This includes my study, *Do Prosody Transfer Models Transfer Prosody?* (Sigurgeirsson and King, 2023), presented in Chapter 5.

Although the primary focus is on PT models, several other reference-based *tasks* are discussed throughout the current part. For example, I cover Global Style Token (GST) (Wang et al., 2018) models in Section 3.3. These are reference-based models which aim to model speaking styles and are featured in a study later on in the thesis. Then, Chapter 6 presents work on *voice-cloning* models which aim to model speaker identity based on limited reference speaker data. These three applications — prosody transfer, speaking style modelling, and voice cloning — differ in the features they aim to control. Although speaking style modelling and voice cloning are fundamentally different from the PT task, it seems that feature entanglement is a persistent issue across diverse applications of reference-conditioning.

3.2 Prosody transfer

The goal of a PT model is to transfer the reference prosodic characteristics to generated speech, while maintaining the identity of a designated target speaker. This process is illustrated in Figure 3.1. Typically, this is accomplished by jointly training a *reference encoder* alongside an acoustic model as first presented by Skerry-Ryan et al. (2018) and subsequently extended in various studies (Akuzawa et al., 2018; Battenberg et al., 2019; Lee and Kim, 2019; Zhang et al., 2019; Klimkov et al., 2019; Karlapati et al., 2020; Zaïdi et al., 2022). The reference encoder models a highly constrained representation of the reference speech and is trained to produce a single fixed-size *reference embedding*. The underlying acoustic model is conditioned on this latent reference embedding to replicate the reference prosody during speech generation.

PT models have demonstrated the ability to capture and reproduce a range of prosodic characteristics from the reference utterance. They can transfer both utterance-level properties, such as the general shape of the F_0 contour and utterance duration, as well as local fine-grained features, such as the intonation of individual words (Skerry-Ryan

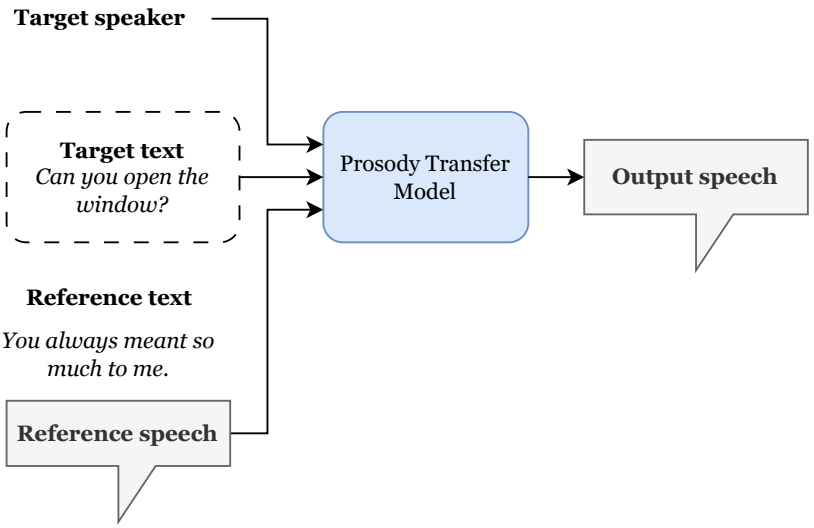


Figure 3.1: In PT, we aim to transfer the prosody from a reference to a synthesised target text in a target speaker's voice. Successful PT involves performing the transfer while preserving the target text and speaker.

et al., 2018). PT models can also, to some degree, produce prosodic variance beyond that demonstrated in the training data (Skerry-Ryan et al., 2018; Zaïdi et al., 2022; Karlapati et al., 2020). Because of this, PT models don't need supervised training to generate expressive speech. Instead, prosodic variation is achieved implicitly by conditioning on an externally provided and appropriately chosen reference utterance. Therefore, the PT approach allows the model to generate expressive speech without explicit training on expression-labelled speech.

To support prosody *transfer*, the reference encoder must construct a robust latent space of prosody. That is, the objective is for samples drawn from this space to produce diverse yet plausible renditions for the given target text and speaker (Skerry-Ryan et al., 2018). During inference, any reference can be used: it can be spoken by the target-speaker or some other speaker (*cross-speaker PT*), and the reference can contain the same verbal information or not (*cross-text PT*).

The prosody transfer terminology

The terms *cross-text* and *cross-speaker* are frequently used in the literature for describing inference conditions where either the reference text or speaker differs from the target. For clarity, especially in Chapter 5, I refer to these as *different-text* and *different-speaker* instead.

The central challenge in PT is to model a prosody representation that generalises across speakers and texts, while preserving the target speaker and intended verbal content. In practice, a PT model can be conditioned to transfer the prosody from any reference to any target text or speaker, and several PT models claim to enable this. For this to work reliably, the reference embeddings must be invariant of the source-text and source-speaker to apply to any text and speaker.

3.2.1 Perspectives on prosody in PT

Evaluating prosody transfer requires a common understanding of what prosody actually is. Traditional definitions of prosody can be categorised into views of prosody based on either the pragmatic function of prosody or its form (Wagner and Watson, 2010). The pragmatic view emphasises the role prosody plays in communication—for either augmentive or affective purposes (Taylor, 2009, p. 123-126)—independent of the lexical items used (Ladd, 2008). The formal view considers prosody in terms of the phonetic and phonological realisation of the utterance, focusing on *suprasegmental* features such as intonation, duration and syllable structure (Ohala, 1975).

The Sauce of the Sentence

Personally, I am very fond of this analogy provided by Anne Cutler and Steve Isard:

> *"Prosody is the sauce of the sentence— it adds to, enhances or subtly changes the flavour of the original. And like a good sauce, the realization of a sentence's prosodic structure is a blend of different ingredients none of which can be separately identified in the final product."* (Cutler and Isard, 1980)

In contrast, views of prosody within the PT literature are often less clearly defined than those found in linguistic, phonological and phonetic literature. Rather than transferring individual prosodic features or events, works in the PT literature emphasise the holistic transfer of prosody from the reference to the target. In the foundational work on PT (Skerry-Ryan et al., 2018), they deploy a subtractive definition of prosody, framing it as the signal variation left over after accounting for variation due to phonetic content, speaker identity and channel effects. This view is technically compatible with the formal view of prosody, as it directly treats prosody as a component in the acoustic signal. However, this simplified, residual view assumes a clean decomposition of the signal into these four components and does, for example, not account for the fact that individ-

ual speakers vary in how they use prosody (e.g., Cole and Shattuck-Hufnagel, 2011). Therefore, a particular acoustic cue—as evidence of a particular prosodic feature—may only apply to the specific speaker under analysis (Cole and Shattuck-Hufnagel, 2016, p. 3).

3.2.2 The foundational architecture

The PT method was first introduced by Skerry-Ryan et al. (2018), who proposed a novel *reference encoder* architecture jointly trained with a **Tacotron**-based (Wang et al., 2017) model. The joint network is trained like most other end-to-end TTS models: given the target text, the network is trained to minimise the loss of the mel-spectrogram. However, the model used in Skerry-Ryan et al. (2018) makes this prediction additionally dependent on the encoded reference utterance. The reference encoder is jointly trained to minimise the loss of the mel-spectrogram, without any additional training objectives or loss terms to guide its training. The reference embedding—a fixed-size embedding generated by the reference encoder from the ground-truth mel-spectrogram—is concatenated to the latent encoder sequence produced by the **Tacotron** text encoder as shown in Figure 3.2. To support multi-speaker synthesis and different-speaker PT, a speaker embedding—generated by a separate speaker encoder—is additionally concatenated to the latent sequence.

Ideally, the reference encoder captures only characteristics relevant to the reference prosody, while information about the target speaker identity is provided separately from the speaker encoder. The architecture of the reference encoder is relatively simple: convolutional layers down-sample the input mel-spectrogram, and a recurrent layer is used to generate a summarisation of the spectrogram in a single fixed-size embedding. Since the joint PT model is trained to only minimise the target mel-spectrogram loss, we can view the training objective as minimising the reconstruction loss of the reference. That is, the model learns to reconstruct the reference mel-spectrogram, which is identical to the target mel-spectrogram, with the speaker label and target text as conditioning inputs.

Objective evaluations in Skerry-Ryan et al. (2018) demonstrate that the proposed approach can capture and transfer acoustic correlates of prosody from the reference, even if spoken by an unseen speaker or containing verbal content unrelated to the target text. The proposed model also generates speech that is perceptually more similar to the reference than the baseline, which is no surprise since the model has direct access to the

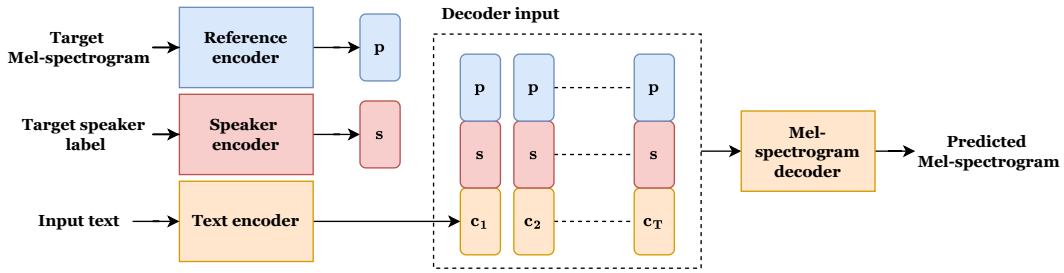


Figure 3.2: High-level overview of model proposed in Skerry-Ryan et al. (2018). The reference and speaker embeddings are repeated across time and concatenated to the encoded text sequence before being passed to the decoder.

reference, while the baseline does not. In Skerry-Ryan et al. (2018), conditioning on a speaker label different from the reference speaker can result in speech that is still identifiable as the target speaker, but with prosodic characteristics transferred from the reference.

3.2.3 Evaluation strategies

Following the approach used in Skerry-Ryan et al. (2018), PT models are commonly evaluated using a combination of objective and subjective metrics (e.g., Battenberg et al., 2019; Klimkov et al., 2019; Lee and Kim, 2019).

3.2.3.1 Objective metrics

Based on the residual definition of prosody, acoustic similarity between the synthesised target and the reference is taken as an objective indication that the model can transfer prosody. Objective metrics are conventionally used only to evaluate the training condition, specifically when the target matches the reference. Several metrics have been proposed for measuring this similarity, but Skerry-Ryan et al. (2018) used Mel-Cepstral Distortion (MCD) (Kubichek, 1993), F_0 Frame Error (FFE) (Chu and Alwan, 2009), Gross Pitch Error (GPE) (Nakatani et al., 2008) and Voicing Decision Error (VDE) (Nakatani et al., 2008). These metrics assess acoustic differences between two signals and assume that the signals are of equal duration. Since this condition is typically not met in PT, Skerry-Ryan et al. (2018) address the mismatch by padding the shorter signal up to the duration of the longer one[1].

[1] While using Dynamic Time Warping (DTW) to align the two signals would have been more appropriate.

Mel-Cepstral Distortion was originally proposed as a measure for general speech quality, replacing Cepstral Distortion (CD) by grounding the evaluation on the perceptually motivated mel-frequency scale (Kubichek, 1993). MCD computes the square of the summed difference between reference and predicted Mel-Frequency Cepstral Coefficients (MFCCs). Assuming K MFCCs, the MCD measure for frame t from the reference signal x and the prediction y is given as:

$$\mathrm{MCD}_t(x,y) = \sqrt{\sum_{k=i}^{K} [MC_x(t,k) - MC_y(t,k)]^2}$$

where $MC(t,k)$ denotes the kth Mel-cepstral coefficient for frame t. The 0th coefficient is typically omitted as it primarily reflects signal energy rather than speech quality. Skerry-Ryan et al. (2018) used 13 MFCCs and computed the average distortion over all frames, T in total, resulting in an utterance-level measure:

$$\mathrm{MCD}(x,y) = \frac{1}{T}\sum_{t=1}^{T} \sqrt{\sum_{k=1}^{13} [MC_x(t,k) - MC_y(t,k)]^2}$$

This measure computes the temporal average mean-squared MFCC difference between the reference and target. It does, therefore, not encode pitch-related information, like mean F_0 or intonation patterns. However, it computes the overall spectral envelope difference between the reference and target, which may reflect other suprasegmental features of prosody.

Gross Pitch Error can complement MCD as it estimates the overall difference in F_0 between the reference and the target. GPE estimates the proportion of voiced target frames that deviate from the reference F_0 at the corresponding frame, given an error threshold (Nakatani et al., 2008). In Skerry-Ryan et al. (2018), a pitch error is defined as any frame where the target F_0 falls outside a $\pm 20\%$ range of the reference F_0.

While GPE focuses on pitch error in voiced regions, it does not account for whether voicing itself is correctly reproduced in the output. To address this, Voicing Decision Error is used to evaluate differences in voicing between the reference and the synthesised output. VDE is the proportion of incorrectly voiced frames in the output, compared to the reference (Nakatani et al., 2008). This metric may reflect broader prosodic differences—such as intonation, rhythm, and stress—as VDE captures variations in both the number and length of voiced segments.

These two metrics, GPE and VDE, are combined in the unified F_0 Frame Error metric. FFE computes the fraction of output frames that contain either a pitch error or a voicing decision error (Nakatani et al., 2008):

$$\text{FFE}(x,y) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left[|p_x(t) - p_y(t)| > 0.2p_x(t) \wedge v_x(t) \wedge v_y(t)\right] + \mathbf{1}[v_x(t) \neq v_y(t)]$$

where p_x and p_y are the estimated F_0 contours for the reference and the output, $v(t)$ indicates a voicing decision at frame t and $\mathbf{1}$ is the indicator function.

These objective metrics are only reliable indicators of prosody transfer under the same-text, same-speaker condition. For instance, FFE does not account for the natural difference in speakers' mean F_0, often resulting in inflated error scores in different-speaker PT. Likewise, both MCD and FFE are sensitive to misalignments in timing. As a result, these metrics are not well suited for evaluating different-text PT samples where such an alignment cannot be assumed. Battenberg et al. (2019) used DTW to find the minimal MCD between the two samples. While this may somewhat mitigate alignment issues, objective evaluations are still limited and typically only serve as developmental indicators of model performance. Objective evaluations are, therefore, supplemented with an evaluation of perceived prosody transfer.

3.2.3.2 Subjective evaluation

To evaluate the model's ability to transfer prosody, Skerry-Ryan et al. (2018) introduces a *anchored prosody side-by-side* AXY discrimination task. Evaluators were presented with three samples: a natural reference sample (A) and two competing synthetic stimuli (X and Y). In this experimental design, A is the reference used to generate the PT sample Y. In Skerry-Ryan et al. (2018), the competing stimulus, X, is generated using a baseline TTS model without a reference encoder. Evaluators were then asked to indicate which stimulus is more prosodically similar to A on a 7-point scale, ranging from "X is much closer" to "Y is much closer". Listeners were instructed to pay attention to four prosodic features when making their judgements: (1) the pitch, and the intonation throughout the utterance; (2) stress, both word and syllable stress reflected either in loudness or change in pitch; (3) speaking rate, and how it may change over time; and (4) the length of pauses.

An alternative to this design, which allows for comparing the transfer abilities of multiple models, is a MUSHRA-like design like the one used by Zaïdi et al. (2022). In a

standard Multi Stimulus with Hidden Reference and Anchor (MUSHRA) design, participants rate several stimuli on a scale from 0 to 100 after listening to a reference. To calibrate the scale, listeners also rate a hidden reference, which should reflect the optimal capability of the model, and one or more intentionally degraded anchors. However, in the context of cross-text PT, a true hidden reference is typically not available and was, therefore, omitted in Zaïdi et al. (2022). But, they did include an anchor stimulus, where they artificially degraded a synthesised sample by randomising phone durations and flattening the F_0 contour. These modified samples are unlikely to resemble the prosody of the references they used and are, therefore, expected to receive the lowest rating among all evaluated samples.

Prosody transfer involves transferring the prosody from the reference while preserving the target speaker's identity. But, the discriminative task proposed in Skerry-Ryan et al. (2018) and the MUSHRA-like design used in Zaïdi et al. (2022) do not assess this preservation. So, an additional speaker identity evaluation task is often performed as well (e.g., Zaïdi et al., 2022). Additionally, to evaluate potential naturalness degradation—particularly for cross-speaker or cross-text synthesis—a standard Mean Opinion Score (MOS) survey is commonly carried out (e.g., Battenberg et al., 2019; Zaïdi et al., 2022).

3.2.4 Challenges

The PT task, as first presented in Skerry-Ryan et al. (2018), demonstrated a new potential way of controlling TTS models. A reference utterance can be used to guide prosody generation, but Skerry-Ryan et al. (2018) also discuss two main problems they encountered. First, conditioning the model on a reference with a vastly different phrase structure to the target results in undesirable prosody transfer and reduced naturalness (Skerry-Ryan et al., 2018). This issue is true in general, but different approaches to PT make different assumptions about what the results of such a transfer should be. In Skerry-Ryan et al. (2018), the poor different-text performance is attributed to the model's failure to disentangle prosody from other aspects of the reference utterance. This issue, referred to as *acoustic feature entanglement*, results in the model failing to render the target text appropriately. For a representation of prosody to be truly transferable, it must encode the reference prosody in a way that can be flexibly and appropriately applied to any target text.

The second issue identified by Skerry-Ryan et al. (2018) concerns the perceived identity of the synthesised speaker. Although the model can transfer prosody from one speaker to another, it does so in a *pitch-absolute manner*. That is to say, the absolute pitch replacement compromises the identity of the target speaker with the reference speaker. So, even if other aspects of the speaker's identity, such as perceived gender and timbre, may be preserved, the overall speaker identity is entangled with features modelled by the reference encoder. This results in the target speaker's identity being mixed with features related to the reference speaker. This particular phenomenon is often described as *source-speaker leakage* (Karlapati et al., 2020).

PT models are invariably trained in the *same-speaker*, *same-text* setting, where the reference exactly matches the target. Yet, PT models are used and evaluated in different-text and different-speaker settings. It is under these conditions that entanglement and source-speaker leakage become prominent. Because of how PT models are trained, they fail to generalise to different-text and different-speaker conditions effectively: the reference must be similar to the target text and speaker to avoid entanglement issues. Finding a suitable reference utterance is non-trivial as a result. Even though PT models hypothetically allow for conditioning on references from unseen speakers, conditioning any reference-based model on references atypical of the training corpus may negatively impact naturalness (Wang et al., 2018; Hsu et al., 2019), further restricting the pool of plausible references. Several subsequent approaches have attempted to address the feature entanglement problem through modelling techniques (Battenberg et al., 2019; Karlapati et al., 2020; Zhang et al., 2019; Zaïdi et al., 2022). However, none directly address the discrepancy between how these models are trained and how they are used during inference.

3.2.5 Extended architectures

Most subsequent prosody transfer methods build on the framework introduced by Skerry-Ryan et al. (2018), typically comprising a core acoustic model—commonly **Tacotron** (Wang et al., 2017)—and a reference encoder similar to that of the original design. Where these approaches primarily differ lies in the choice of conditioning inputs for the reference encoder during training, and in the strategies employed to integrate this information into the acoustic model. The differences usually reflect efforts to find a balance between the perceived prosody transfer quality and minimising the entanglement of prosody with other acoustic features in the latent representation of prosody (Battenberg et al., 2019).

3.2.5.1 Representational capacity

The representational *capacity*, or the amount of information that can be transferred from the reference to the output, is a crucial component of the reference embeddings. There is a trade-off between high-capacity and low-capacity embeddings (Battenberg et al., 2019). If the dimensionality of the reference embedding is highly limited, the capacity might become too low to support prosody transfer. On the other hand, if the capacity is too high, the network would naturally learn to transfer all the information from the reference, since this would minimise the target mel-spectrogram loss, as explained in Skerry-Ryan et al. (2018). Therefore, too high a capacity amplifies feature entanglement issues under cross-text and cross-speaker conditions.

Low-capacity representations (Wang et al., 2018; Akuzawa et al., 2018; Zhang et al., 2019, for example) are effectively bottlenecked by their capacity. They cannot model the same level of detail as demonstrated in Skerry-Ryan et al. (2018), which is required for the PT task. Instead, low-capacity models are deployed for tasks like modelling speaking style or emotions. Representations of higher capacity than that in Skerry-Ryan et al. (2018), such as Lee and Kim (2019) and Karlapati et al. (2020), model a highly reference-specific representation. Such high-capacity representations are used for tasks such as *voice-puppetry* where we expect the lexical content of the reference to be highly similar to the target text. Therefore, in PT, the aim is to choose the appropriate representational capacity: the capacity must be sufficiently large to enable prosody transfer. At the same time, it must be sufficiently small to force the reference encoder to model only features that are not determined by other inputs, such as the target speaker and text.

3.2.5.2 Model-based strategies for disentanglement

Finding the right representational capacity may not be sufficient to mitigate source-speaker leakage and feature entanglement. So, several approaches seek to address these challenges through alternative model-based strategies.

In Battenberg et al. (2019), the modelling of the reference embedding is made conditionally dependent not only on the reference mel-spectrogram but also on the reference speaker and text. They argue that this improves performance in different-speaker or different-text PT, as the model is explicitly informed of the mismatch between the reference and the intended output. A similar strategy is employed by Karlapati et al.

(2020), where the reference embedding generation is conditional on the source speaker. However, they observe that the model learns to depend on the reference embedding for modelling speaker identity when the embedding capacity is sufficiently large to support PT. Without additional mitigations, this leads to source-speaker leakage. To address this, they propose using a temporal *bottleneck encoder* (Qian et al., 2019). This encoder limits the influence the reference encoder has on the decoder along the time axis. This limitation, therefore, forces the underlying acoustic model to rely on other signals for learning the speaker identity. Others have proposed speaker-normalising the reference embeddings to alleviate source-speaker leakage (Lee and Kim, 2019). However, this requires sufficient source data from the reference speaker to perform effective normalisation.

Daft-Exprt is a parallel PT model architecture proposed for expressive different-text and different-speaker PT (Zaïdi et al., 2022). **Daft-Exprt** employs several modelling techniques to address feature entanglement. The model generates a fixed-size representation of prosody, similarly to Skerry-Ryan et al. (2018), to condition speech generation. The information capacity of the modelled representation of prosody is also comparable to that of Skerry-Ryan et al. (2018).

A daft name

Daft-Exprt is a prominent model in the experiments conducted in Chapter 5 and is, therefore, formatted in bold throughout this thesis. This curious model name is derived from the core contribution in that work: **D**eep **af**fine **t**ransformations for **E**xpressive **p**rosody **t**ransfer.

But **Daft-Exprt** is different from the other typical PT models in several ways. **Daft-Exprt** is based on **FastSpeech 2** (Ren et al., 2020) while most other work on PT uses a **Tacotron** architecture (for example Skerry-Ryan et al., 2018; Zhang et al., 2019; Battenberg et al., 2019; Lee and Kim, 2019; Klimkov et al., 2019). Figure 3.3 shows an overview of how **Daft-Exprt** is trained. The acoustic model is similar to **FastSpeech 2**, consisting of a phoneme encoder, a *local prosody predictor* and a frame decoder. Similar to **FastSpeech 2**, **Daft-Exprt** predicts phone-level F_0, energy and duration in the local prosody predictor. But different from **FastSpeech 2**, they employ Gaussian upsampling (Shen et al., 2020) instead of a length regulator for upsampling the phone-level sequence before decoding the mel-spectrogram. The roles of these two different modules are the same, but Zaïdi et al. (2022) claim that using Gaussian upsampling instead of a length regulator improves naturalness.

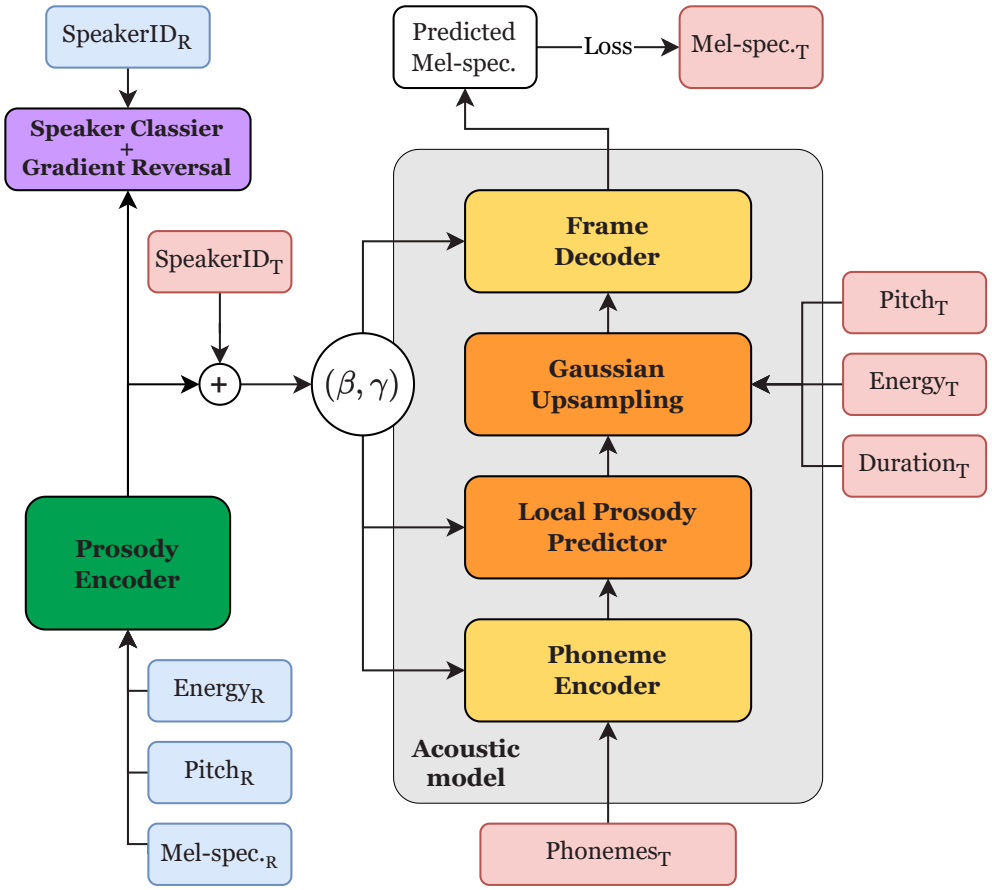


Figure 3.3: **Daft-Exprt** training involves a target utterance (information shown in red and indicated with 'T') and a reference utterance (information shown in blue, indicated with 'R'). **Daft-Exprt** conditions the acoustic model, at three locations, using the predicted FiLM parameters (β and γ).

The reference encoder, referred to as a *prosody encoder* in Zaïdi et al. (2022), predicts an embedding from the reference mel-spectrogram. Optionally, the extracted utterance F_0 and energy contours can be used as additional inputs. **Daft-Exprt** also employs a modelling technique to address source-speaker leakage specifically. The model jointly trains a speaker classifier with gradient reversal (Ganin and Lempitsky, 2015) to minimise the influence of the source-speaker on speech generation. After predicting the reference embedding, this speaker classifier tries to predict the *reference* speaker from that reference embedding, with gradient reversal discouraging the embedding from containing reference-speaker identity.

After this, the *target* speaker identity is separately embedded and summed to the reference embedding to indicate which voice to synthesise. This combined representation is used as the conditioning input for the acoustic model, supporting both PT and speaker modelling. In other work on PT, the reference embedding is typically incorporated into

the acoustic model using concatenative or additive conditioning (Skerry-Ryan et al., 2018; Lee and Kim, 2019, for example). **Daft-Exprt**, instead, uses Feature-wise Linear Modulation (FiLM) (Perez et al., 2018) to incorporate the reference prosody representation. FiLM conditioning is a general-purpose conditioning method that allows for influencing the output of a neural network. Conditioning is achieved by applying an affine transformation to the model's intermediate features in a chosen location. Given some input $\mathbf{x}$, FiLM learns functions to predict scaling (γ) and biasing (β) parameter arrays to enable this transformation (Perez et al., 2018). When applied to a model layer that predicts an intermediate sequence $\mathbf{S}$ of shape $[T, D]$, for example, FiLM learns scaling and biasing parameters to transform $\mathbf{S}$ by:

$$\text{FiLM}(\mathbf{S}_{t,d}) = \gamma_d \mathbf{S}_{t,d} + \beta_d \qquad d = 1 \dots D, \quad t = 1 \dots T \tag{3.1}$$

FiLM enables very flexible conditioning, as the transformation can amplify specific activations and selectively threshold others. The functions predicting the γ and β parameters can be learned with a neural network jointly trained with the underlying network. In Zaïdi et al. (2022), these functions are learned by simple feed-forward networks, taking the learned reference embedding as part of the input $\mathbf{x}$. Compared to concatenation or additive conditioning, FiLM allows for independent conditioning of several parts of the model from the same representation. To enable multi-location conditioning, the size of the learned γ and β parameter arrays is expanded, and then subsets of those parameters are separately applied at different locations in the model. In Zaïdi et al. (2022), there are FiLM-conditioning layers at three strategic locations in the acoustic model: the phoneme encoder, the *local prosody predictor*, and the frame decoder. The reference embedding thus influences the learned phoneme representation, predicted F_0, energy, duration, and mel-spectrogram.

The combined model is trained to minimise a loss comprising four terms:

$$\mathcal{L} = \mathcal{L}_e + \lambda_f \mathcal{L}_f - \lambda_a \mathcal{L}_a + \lambda_r \mathcal{L}_r \tag{3.2}$$

where $\mathcal{L}_f$ regularizes the scale of (γ, β) FiLM parameters as in Oreshkin et al. (2018) and $\mathcal{L}_r$ corresponds to weight decay. $\mathcal{L}_a$ is the speaker classifier loss, the $(-)$ indicating gradient reversal. Duration, energy, F_0, and mel-spectrogram losses are combined in $\mathcal{L}_e$, which is the same as the standard loss for the core acoustic model (Ren et al., 2020) with the addition of Mean Absolute Error (MAE) of the predicted mel-spectrogram.

It is found that when the influence of the speaker classifier—λ_a in Equation 3.2—is increased, the model better preserves target-speaker identity but at the cost of perceived prosody transfer.

Zaïdi et al. (2022) report strong results in a different-text, different-speaker PT task, while maintaining comparatively high perceived naturalness. **Daft-Exprt** is compared to a GST model (Wang et al., 2018), a model that employs a Variational Autoencoder (VAE) (Zhang et al., 2019), and Flowtron (Valle et al., 2020b) and is rated significantly better than each one in the PT task.

3.3 Global style tokens

One issue with PT models is that they always require a reference for conditioning. This requirement can be a limitation: many use cases rely on the TTS model to generate speech in general while only guiding it when necessary. GST models (Wang et al., 2018) are extensions of PT models, which address this limitation. GST models incorporate a finite bank of *global style tokens*. These are randomly initialised embeddings which are jointly trained with the acoustic model (**Tacotron** in Wang et al. (2018)). The number of tokens employed varies between works. The original work uses 10 tokens, whereas (Prateek et al., 2019) uses only two tokens, and (Wu et al., 2019) uses 256 tokens. Like PT models, GST models condition speech generation on a reference embedding of the target mel-spectrogram during training. Using attention, the model determines similarity between the reference embedding (or *style embedding*) and all GSTs. A weighted sum of the global style tokens, where weights correspond to similarity to the reference, is then used to condition the model.

In this training regime, the model learns to encode training corpus variation into a finite list of style tokens. When trained on expressive speech, these tokens demonstrate the capacity to model speaking styles (Wang et al., 2018; Valle et al., 2020a; Hsu et al., 2019) and emotions (Wu et al., 2019; Kwon et al., 2019; Sorin et al., 2020). The flexibility of this approach allows for several different types of inference modes: (1) using a reference to find a corresponding mixture of style tokens; (2) manual selection of a token, or a mixture of tokens; (3) scaling of token weights. So during inference, a reference is optional, as the model can employ a combination of style tokens instead.

Although GST models can achieve inference without a ground-truth reference, it is non-trivial to find the right combination of tokens to achieve a desired target style

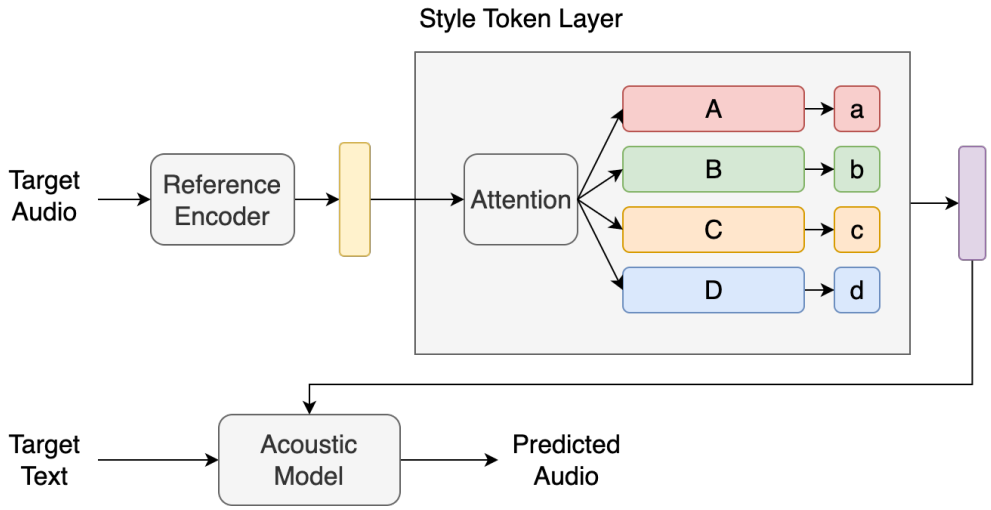


Figure 3.4: The style token layer is queried using the reference embedding (yellow) as a key. Using attention, the similarity (a-d) between all GSTs (A-D) and the reference embedding is determined. Using these similarities as weights, a single weighted-sum (purple) of the style tokens, $aA + bB + cC + dD$, conditions the acoustic model.

(Wang et al., 2018; Valle et al., 2020a). A GST model is trained to capture all acoustic variation from the training corpus. Speaker identity, residual variance, and noise are thus also modelled by style tokens (Wang et al., 2018). GST models are, therefore, also susceptible to feature entanglement and conditioning on out-of-distribution references can lead to unstable results (Hsu et al., 2019). Methods have later been proposed to address feature entanglement (Wu et al., 2019; Kwon et al., 2019). The style tokens themselves are also entangled, since there is no guarantee that different tokens model perceptually distinct vocal behaviours in a separable manner. Some works employ an additional loss term to force such separation (Wu et al., 2019) while others have used labelled data to determine what each token represents (Kwon et al., 2019). Others condition speech generation on speaker labels to encourage separation between style tokens and speaker identity modelling (Valle et al., 2020a; Battenberg et al., 2019).

3.4 Variational autoencoders

The reference-based models discussed so far are based on non-variational architectures (e.g., Skerry-Ryan et al., 2018; Wang et al., 2018). That is, the conditioning of a particular reference utterance always yields the same reference embedding, thus the same speech output. Many have employed VAEs (Kingma and Welling, 2013) to enable variational reference-based generation (e.g., Akuzawa et al., 2018; Zhang et al.,

2019; Battenberg et al., 2019; Zhang et al., 2019; Sun et al., 2020; Karlapati et al., 2020). The studies presented later in this part do not employ variational architectures; however, they are discussed here due to their relevance to the topic.

VAEs differ from conventional autoencoders in that they are non-deterministic. Instead of constructing an embedding in the latent space, the encoder models a distribution over the latent space, from which a sample is drawn at generation. Given the sample, the decoder performs the reconstruction (Kingma and Welling, 2013). Like other auto encoders, VAEs will learn to model *any* speech signal variation, not accounted for by other means, that minimises the reconstruction loss of the mel-spectrogram. Conditioning VAE generation (Sohn et al., 2015) can influence what sort of features are modelled by the VAE. For example, Kenter et al. (2019) conditions prosody generation on linguistic features to predict word-level prosody features. The VAE architecture itself does not provide strong support for modelling disentangled features. Still, extended implementations of this architecture, e.g. β-VAE (Higgins et al., 2017), have been proposed to support more effective disentanglement of learned features for TTS (e.g., Zhang et al., 2019). From a controllability perspective, VAEs also offers the benefit of *smooth interpolation* between latent samples, yielding a gradual transition between the characteristics of the two samples (e.g., Akuzawa et al., 2018; Zhang et al., 2019).

3.5 Focus of presented work

The following three content chapters include research on reference-based models, focusing primarily on PT. First, a pilot study on the PT task is presented in Chapter 4. Then, two published papers follow; one researching training strategies for PT (Chapter 5) while the other focuses on voice-cloning (Chapter 6). Central to all these chapters is the fundamental challenge inherent to reference-based control methods: feature-entanglement.

Chapter 4

Pilot study: a parallel prosody transfer model

Before **Daft-Exprt** was proposed by Zaïdi et al. (2022), I conducted a pilot study exploring similar strategies to mitigate source-speaker leakage in Prosody Transfer (PT). Like in Zaïdi et al. (2022), the pilot study explored replacing **Tacotron** (Wang et al., 2017) as the core acoustic model in PT, motivated by several limitations. As an autoregressive architecture, **Tacotron** has a high computational cost during both training and inference. **Tacotron** also suffers from alignment issues, as previously covered in Section 2.3.2.2. So, I proposed a PT model based directly on the **FastSpeech 2** (Ren et al., 2020) architecture instead. **FastSpeech 2** predicts the entire mel-spectrogram in parallel and replaces attention with a duration predictor to perform the text-to-mel-spectrogram alignment. Together, these result in a more efficient and robust Text-To-Speech (TTS) pipeline. Using models like **FastSpeech 2**—which explicitly model acoustic correlates of prosody—for the PT task may also benefit controllability, interpretability, and analysis of the model, later discussed in Chapter 9 and Chapter 10.

Initially, the primary study objective was to develop the first fully parallel PT model and, then, train it on an Icelandic speech corpus. However, the focus shifted to investigating different conditioning inputs for the PT reference encoder to address source-speaker leakage. In Skerry-Ryan et al. (2018), the reference encoder is only conditioned on the reference mel-spectrogram. However, other features can be used to guide prosody prediction, and not all information encoded in the mel-spectrogram is informative of the reference prosody. I conducted experiments using F_0, as a primary acoustic correlate of prosody, as a reference encoder input instead of the mel-spectrogram.

Icelandic as a source language

I approached this first research objective during my PhD, as an Icelander and someone with experience in speech technology in Iceland. Icelandic is a low-resource language[a], and I enjoy contributing to the body of Icelandic TTS research. It is worth noting, however, that the experiments I conducted are not language-specific; the findings extend to the broader application and development of PT models.

[a] Although its resource status has improved rapidly in recent years (Steingrímsson et al., 2024, p. 19).

4.1 Proposed model and data

4.1.1 Reference encoder

To realise PT with **FastSpeech 2**, it was jointly trained with a reference encoder and a speaker identity encoder. The reference encoder draws on the one proposed by Skerry-Ryan et al. (2018). It consists of 6 convolutional layers that down-sample the input mel-spectrogram in both dimensions. This representation is then passed to a Gated Recurrent Unit (GRU) (Cho et al., 2014) layer, and the last hidden emission is taken as the summarising reference embedding. Based on the prior work, $d_P = 128$ was chosen as the dimensionality of the embedding modelled by the reference encoded. I experimented with alternative conditioning inputs for the reference encoder. Specifically, quantised normalised representation of F_0 instead of the mel-spectrogram. After extracting the F_0 contour, it was discretised into 80 speaker-normalised bins, including a bin for unvoiced segments. Because of how the reference F_0 contour was encoded, it matched the expected reference mel-spectrogram dimensionality and, therefore, no changes to the reference encoder architecture were required.

4.1.2 Speaker encoder

I aimed to evaluate the capabilities of the proposed model under typical PT inference conditions, where the target text or speaker may differ from the content of the reference. Therefore, the baseline model would have to support multi-speaker training and generation. **FastSpeech 2** is not a multi-speaker TTS model, so methods from Gibiansky et al. (2017) to make the model's speech generation speaker-dependent were adopted. In Gibiansky et al. (2017), a speaker encoder network conditioned on a one-

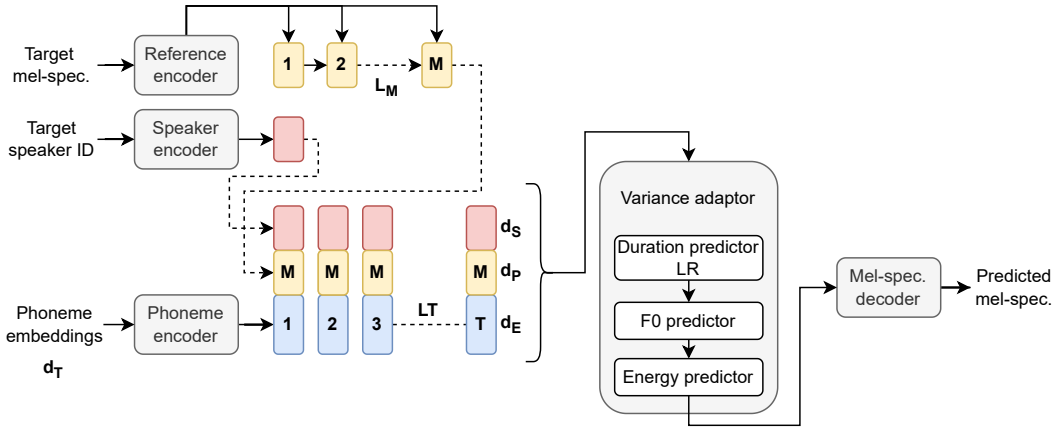


Figure 4.1: The proposed model consists of three encoders: speaker, reference (or prosody), and text. **D**s and **L**s are used to indicate dimensionalities and sequence lengths, respectively. The speaker and reference embeddings are concatenated to the text encoder output and then passed to the variance adaptor. Finally, a decoder predicts the mel-spectrogram.

hot speaker label is trained jointly with a *Deep-Voice* TTS network (Arik et al., 2017) to minimise the loss of the predicted mel-spectrogram. The speaker encoder generates a fixed-size embedding for each speaker in the training corpus, and the acoustic model is conditioned on the target speaker's embedding during training. Another approach would be to train multiple speaker-dependent acoustic models. But, incorporating a speaker encoder enables parameter sharing between voices, reducing overall computation. The original work demonstrates high-quality multi-speaker performance for hundreds of speakers using only[1] 30 minutes of source-speaker data, per speaker (Gibiansky et al., 2017). I used a single feed-forward layer to project the speaker IDs to fixed-size multidimensional embeddings. Initial analysis suggested that 64 dimensions ($d_S = 64$) were sufficient for the corpus and model used in the study.

4.1.3 Acoustic model and training

In Skerry-Ryan et al. (2018), the reference and speaker embeddings are concatenated to the text-encoder output to condition speech generation. Where the reference and speaker embeddings are introduced in the acoustic model determines which parts of the model are influenced by the information these embeddings encode. For example, the speaker embedding could be introduced just before mel-spectrogram decoding for tighter timbre control (Qian et al., 2020). Importantly, the variance adaptor predicts

[1]This was considered impressive at the time.

features reflecting both speaker characteristics and prosody. So, both embeddings are concatenated to the text encoder output across all time steps as shown in Figure 4.1. The **FastSpeech 2** encoder produces a latent $L_T \times d_E$ sequence, where d_E is the latent encoder dimensionality and L_T is the number of encoder time steps. So, after concatenative conditioning, the intermittent representation dimension becomes $d_E + d_P + d_S$, as shown in Figure 4.1.

The F_0, energy, and duration predictors are trained to minimise Mean Square Error (MSE) loss of corresponding ground truth values. The reference and speaker encoders are jointly trained with the TTS model without further supervision. The whole model is trained to minimise the sum of the variance prediction losses and $L1$ loss of the predicted mel-spectrogram. The Adam optimiser (Kingma and Ba, 2014) is used for all experiments. Gradient clipping (Goodfellow et al., 2016) was employed for norms ≥ 1.0, to mitigate exploding gradients. A WaveGlow (Prenger et al., 2019) vocoder, pretrained on the LJSpeech corpus (Ito and Johnson, 2017), was used for all experiments. An overview of other hyperparameters is given in Table A.1 in Appendix A.

4.1.4 Data

The proposed model was trained on Talrómur, a large Icelandic TTS corpus (Sigurgeirsson et al., 2021). The corpus consists of four female and four male speakers, comprising approximately 120,000 recordings, resulting in roughly 213 hours of speech. The majority of these recordings are between 5 and 8 seconds long. The text prompts used in the corpus are sourced from various books and newspapers, and voice talents read the prompts in a neutral narration style.

Talrómur

Talrómur [tʰalroumʏr] is an apt name for a speech corpus as *Tal* means *speech* while *rómur* means *voice*. Names of Icelandic speech resources conventionally include the *-rómur* suffix.

An overview of all corpus speakers is given in Table 4.1. The speakers vary both in pitch range and speaking rate, which play essential roles in communicating speaker identity and prosody. 80% of the corpus data was used for training, split equally between all speakers. A Montreal Forced Aligner (MFA) model (McAuliffe et al., 2017), pretrained on a mixture of the Talrómur voices, was employed for extracting align-

ments to facilitate the training of the duration predictor. Informal experiments show that this aligner is adequate for the task. F_0 contours were estimated using algorithms in the WORLD library (Morise et al., 2016). Because the Talrómur corpus includes no digits, acronyms, or abbreviations, no text normalisation was required. The text prompts were phonemically transcribed to provide phoneme model inputs instead of graphemes. A Sequitur Grapheme-To-Phoneme (G2P) model (Bisani and Ney, 2008), trained on an Icelandic pronunciation dictionary (Nikulásdóttir et al., 2018), was employed for this task.

Speaker ID	# Utterances	Duration (s)	SR (word/s)	Mean F_0 (Hz)
A	9 899	16h:32m:12s	2.34	198.8 ± 22.6
B	12 048	25h:43m:05s	1.73	150.7 ± 24.6
C	13 443	27h:57m:33s	1.89	173.6 ± 24.5
D	12 357	22h:32m:58s	2.24	143.7 ± 28.0
E	20 050	31h:28m:04s	2.94	210.7 ± 27.0
F	19 849	29h:07m:18s	3.26	128.1 ± 12.9
G	16 886	30h:09m:38s	2.39	237.1 ± 20.6
H	17 637	29h:49m:01s	2.60	142.7 ± 23.4

Table 4.1: Summarising information for the eight Talrómur speakers. Speakers A, C, E, and G are female, and the others are male. The reported Speaking Rate (SR) and F_0 are sourced from Sigurgeirsson et al. (2021).

4.2 Model evaluation

No subjective evaluation was carried out in the current work. Instead, F_0 Frame Error (FFE) was used as a primary objective metric for initial evaluation of the quality of prosody transfer. As previously explained in Section 3.2.3.1, metrics like FFE require that the speaker and text spoken in each sample compared are the same. Therefore, speaker-Normalised F_0 Frame Error (NFFE) was used as a secondary metric to account for differences between reference and target speakers. Still, NFFE is insufficient for evaluating different-text PT, so I focused my investigation on the different-speaker, same-text condition instead.

First, a single-speaker baseline (**FastSpeech 2**) was compared with a single-speaker version of the proposed PT model. Then, a multi-speaker baseline was compared with multi-speaker PT models that differed in conditioning inputs. Objective results for all models are summarised in Table 4.2.

4.2.1 Conditioning on mel-spectrogram

The baseline and proposed single-speaker models were both trained on **speaker A** from the Talrómur corpus. As seen in Table 4.2, the FFE of the baseline single-speaker model and the proposed model are comparable under the same-speaker condition. The FFE similarity suggests that the proposed model does not transfer additional information from the reference to help reconstruct the target mel-spectrogram. Therefore, this initial model would not be able to transfer prosody. Using a multi-speaker model, the contribution of the reference encoder can be further evaluated under cross-speaker conditions. However, as the results demonstrate, the proposed multi-speaker PT model—trained on mel-spectrogram inputs— is comparable to a baseline multi-speaker model under both same and cross-speaker conditions. This suggests that the addition of the reference encoder to enable PT has not succeeded.

Model	**FFE**		**Normalised FFE**	
	same speaker	diff. speaker	same speaker	diff. speaker
SS baseline	$43.7\% \pm 8.0\%$	n/a	$39.2\% \pm 11.1\%$	n/a
+ ref. enc.	$46.2\% \pm 14.2\%$	$58.3\% \pm 12.5\%$	$48.1\% \pm 9.9\%$	$52.2\% \pm 14.5\%$
MS baseline	$44.2\% \pm 9.6\%$	$62.3\% \pm 16.6\%$	$52.1\% \pm 12.8\%$	$52.9\% \pm 13.8\%$
+ ref. enc.	$42.0\% \pm 8.9\%$	$60.1\% \pm 14.2\%$	$49.8\% \pm 7.4\%$	$53.8\% \pm 10.3\%$
+ norm. F_0	$30.3\% \pm 1.2\%$	$51.6\% \pm 15.6\%$	$36.7\% \pm 1.1\%$	$46.6\% \pm 10.1\%$

Table 4.2: Summary of different model types across single (SS) and multi-speaker (MS) conditions. The FFE values reported for the baseline single-speaker model are between the synthesised output and a corresponding evaluation utterance.

Increasing the reference embedding capacity, d_P, from 128 to 256 would, hypothetically, enable the reference to have more influence on speech generation. But increasing capacity results in the synthesised utterance taking the mean F_0 of the reference speaker, indicating *source-speaker leakage* like in Skerry-Ryan et al. (2018). An additional experiment was conducted where the speaker embedding is introduced just before decoding the mel-spectrogram, like in (Qian et al., 2020), to make the mel-spectrogram prediction more explicitly dependent on the speaker identity. However, this did not resolve the issue of source-speaker leakage.

4.2.2 Conditioning on a normalised representation of fundamental frequency

I evaluated a multi-speaker PT model conditioned on a quantised and normalised F_0 representation instead of the mel-spectrogram to address the source-speaker leakage.

The estimated frame-level F_0 contour is quantised into 80 bins, which were determined per speaker based on the training corpus statistics. This representation results in a sequence of speaker-normalised F_0 one-hot labels. The number of bins, which matched the number of frequency bins used for mel-spectrograms in the current work, was chosen purely based on compatibility with the model. Such a representation hides mean-F_0 information from the reference encoder and could, hypothetically, help reduce source-speaker leakage. However, this representation also hides energy from the encoder. Energy, which is perceived as loudness, can be necessary for prominence production (Tamburini, 2003). So, using only the F_0 contour as a conditioning input for PT may hinder the model in transferring prosodic characteristics that are otherwise not reflected in F_0.

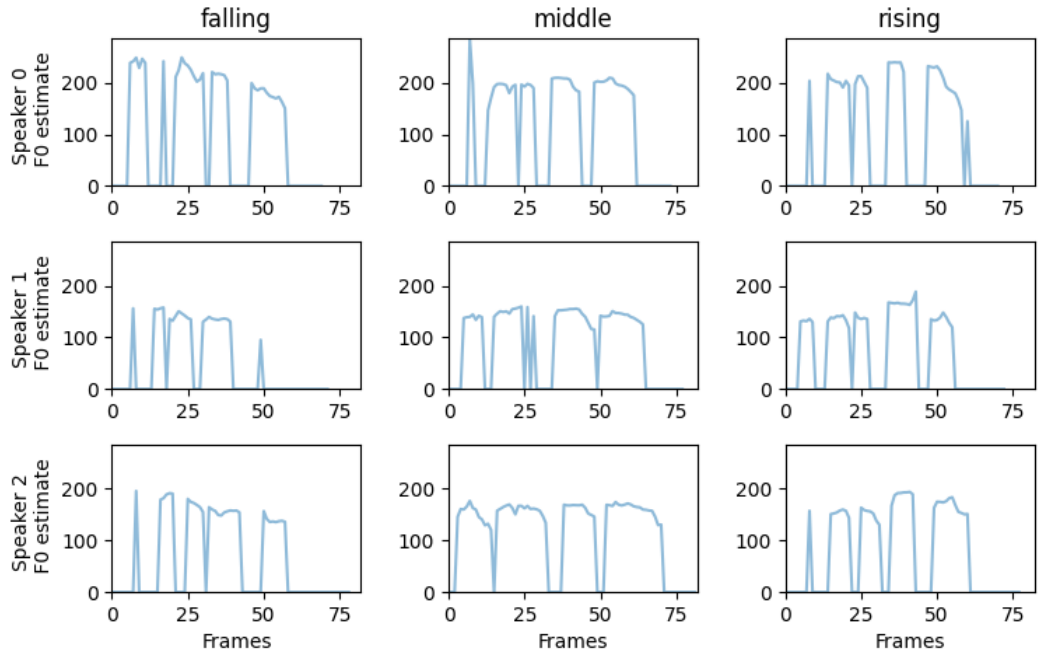


Figure 4.2: F_0 contours (Hz) of synthesis when the model is conditioned on monotonically falling, flat, and rising pitch contours. Results for three corpus speakers demonstrate that the model does, to some degree, transfer utterance-level prosodic information.

The objective results for this model, presented in Table 4.2, are comparably better than the model conditioned on mel-spectrograms. As expected, this form of model conditioning mitigated the problem of source-speaker leakage. Yet, informal listening evaluation suggested that the transferal capabilities of the proposed model were limited to the shape of the F_0 contour. This ability is evidenced in Figure 4.2, where a model is conditioned on three different F_0 contours, using three different speaker labels. When

provided with a rising F_0 contour, the synthesised output mimics the upward trend. The same is true for a falling contour. And conditioning the model on a flat contour yields a flat, monotonic tone.

4.3 Conclusion

Similar to the prior work (Skerry-Ryan et al., 2018), I found that conditioning on the mel-spectrogram resulted in source-speaker leakage. Instead of deploying complex model-based approaches to address this, I suggested changing the conditioning input for the reference encoder instead. Replacing it with a normalised F_0 contour resolves the source-speaker leakage problem while reducing objective error.

This pilot study aimed to propose a candidate model for my future PT-based experiments. The proposed model transfers some limited prosodic information from the reference to the output, mainly reflected in the transfer of the utterance F_0 contour. Yet, the method failed to demonstrate any transfer of duration-related features, like that demonstrated in Skerry-Ryan et al. (2018). **FastSpeech 2**'s duration predictor predicts phone duration, and I hypothesised that this explicitly learned duration control was more salient for reconstructing the mel-spectrogram than the predicted reference embedding. Therefore, alternative modelling techniques would be required to influence duration prediction through the reference embedding.

Nonetheless, the work conducted as part of the case study was timely, as **Daft-Exprt** (Zaïdi et al., 2022) was proposed soon after. This pilot study and their work have many parallels, but they achieve effective reference-conditioning using Feature-wise Linear Modulation (FiLM)-conditioning as previously discussed in Section 3.2.5.2. Similarly, they study different conditioning inputs for the reference encoder, including a quantised representation of F_0. Their model, therefore, became the baseline model for my continuing work on PT and is foundational in the study presented in the next chapter.

Chapter 5

A training strategy for feature disentanglement

Previous work, including the pilot study presented in the previous chapter, highlights that designing a Prosody Transfer (PT) model must consider a trade-off between the quality of PT on one side, and perceived naturalness and preservation of speaker identity on the other. In *Do Prosody Transfer Models Transfer Prosody?* (Sigurgeirsson and King, 2023), I proposed a novel training regime to directly address this trade-off and the entanglement issues covered in Section 3.2.4. However, the findings raise the question whether typical PT models can ever learn prosodic representations that are reliably transferable.

5.1 Research objective

The different PT approaches presented in Chapter 3 assume that the underlying methods can produce transferable representations of prosody. Based on this assumption, I suggested changing how a PT model is trained, rather than changing the model itself, to address the entanglement issues in PT. As previously discussed in Section 3.2.4 (p. 43), PT models are always trained in the *same-speaker*, *same-text* setting. Because of this, they generalise poorly to inference conditions where the target speaker or text does not match the reference text. So, instead of using the target utterance as the reference during training, I employed a reference that is *different*, but that should be informative about the target's prosody. This strategy enables training the model under both different-speaker and different-text settings by selecting a prosodically-informative pair from the training data.

This novel training regime was motivated by two key considerations. First, it replicates the inference conditions, which should lead to better model generalisation. Second, it should encourage the reference encoder to only model features common to both the reference utterance and the target: the prosody. The current work put forward three research questions to pursue this primary research objective:

> **RQ 5-1:** *Can we automatically find prosodically-informative utterances in a speech corpus?*

Key to the proposed method is to train the Text-To-Speech (TTS) model using a reference that is different from the target while being prosodically-informative about it. Two different methods were evaluated for automatically finding such pairs. One method, **text-based**, matches utterances based on lexical content. The other, $\boldsymbol{F_0}$**-based**, employs a F_0 similarity metric to pair utterances. These methods are described in detail in Section 5.2. The following predictions, corresponding to this research question, were made:

> **H5-1a:** *Listeners will rate the reference paired by* ***text-based*** *as more prosodically similar to the target utterance than a randomly selected reference.*

> **H5-1b:** *Listeners will rate the reference paired by* $\boldsymbol{F_0}$***-based*** *as more prosodically similar to the target utterance than a randomly selected reference.*

> **RQ 5-2:** *Does the proposed training regime mitigate the issues of feature entanglement and source-speaker leakage?*

Feature entanglement negatively affects both the preservation of the target speaker identity and perceived naturalness (Skerry-Ryan et al., 2018; Karlapati et al., 2020). This issue arises because PT models fail to disentangle the prosody from other features, relating to speaker identity or the reference lexical content, in the reference utterance. The proposed methods employ references that are prosodically similar to the target utterance, but differ in speaker or lexical content. By changing the training regime, the model is encouraged to only learn what is shared between the reference and the target, therefore mitigating feature entanglement.

Based on these assumptions, the following predictions were made:

H5-2a: *Speech generated by the proposed methods is perceived as more natural under different-text conditions, when compared to* ***Daft-Exprt***

H5-2b: *The proposed methods will better preserve the target speaker identity under different-speaker conditions, when compared to* ***Daft-Exprt***

RQ 5-3: *Can a model trained with different yet prosodically similar references retain the level of prosody transfer demonstrated by a baseline prosody transfer model?*

The proposed training regime is a novel approach for training a prosody transfer model. Therefore, it was not known whether a PT model can learn anything at all given unmatched reference/target pairs. Models trained using the two proposed training methods, **text-based** and $\boldsymbol{F_0}$**-based**, are compared to two additional models: (1) the baseline PT model, **Daft-Exprt** (Zaïdi et al., 2022), where the reference and target are identical during training; and (2) **shuffle**, where the reference is randomly selected from the training corpus. **shuffle** is treated as an *uninformed* model, since a random reference is unlikely to be informative about the target utterance prosody. As such, **shuffle** represented a lower-bound model in the experiments. Section 5.3 gives a detailed description of all models trained as part of this work. The two proposed models were expected to be comparable to **Daft-Exprt** while outperforming **shuffle**:

H5-3a: *The prosodies generated by* ***text-based*** *and* ***Daft-Exprt*** *are equally similar to that of the reference*

H5-3b: *The prosodies generated by* $\boldsymbol{F_0}$***-based*** *and* ***Daft-Exprt*** *are equally similar to that of the reference*

H5-3c: *Compared to* ***Daft-Exprt***, ***text-based***, *and* $\boldsymbol{F_0}$***-based***, *the prosody generated by* ***shuffle*** *is less similar to that of the reference*

5.2 Finding prosodically similar pairs

Two simple ways of selecting a suitable reference for each target utterance were explored to address **RQ 5-1**: **text-based** and $\boldsymbol{F_0}$**-based**.

5.2.1 Text-based method

There is a correlation between the verbal and prosodic components of speech, but predicting this correlation is not straightforward. Different renditions of the same text are, however, likely to have a similar prosodic structure. Therefore, I considered the case where the reference utterance is of the same text as the target utterance but read by a different speaker than the target speaker. Of course, one speaker's interpretation of the text will not be identical to another's, leading to prosodic differences. But, it was assumed that the reference would still be informative for the PT model and help it predict the target prosody.

The *Parallel Audiobook Corpus* (Ribeiro et al., 2018) was used to create the (reference, target) pairs for **text-based**. The corpus was created from LibriVox and comprises over 115 hours of English speech from 49 (22 female and 27 male) North American speakers. All utterances in the corpus are parallel: every text prompt is read by multiple speakers.

5.2.2 F_0-based method

Text-based employs references spoken by a speaker different from the target speaker. The **text-based** strategy should, therefore, discourage the model from learning reference speaker characteristics, thus mitigating source-speaker leakage. But the references selected by **text-based** still match the target lexical content. A selection method based on F_0 similarity was devised to compare with **text-based**. F_0 was chosen since it is the principal acoustic correlate of prosody, influenced by both local and supra-segmental prosodic characteristics. The F_0 similarity-metric was based on Dynamic Time Warping (DTW), using Euclidean distance as in Rilliard et al. (2011). DTW is used to align sequences of per-phone speaker-normalised log-F_0 to select the closest reference for each utterance in the training corpus. This search was limited to utterances that differ by no more than $\pm 15\%$ in phone sequence length.

The parallel audiobook corpus was also used for this training method. After finding the most similar reference for each utterance, any pair with a DTW distance greater than

one standard deviation, based on the mean DTW distance across the entire corpus, was eliminated. The filtering resulted in approximately 55,000 utterance pairs.

5.3 Model training

Answering **RQ 5-3** requires comparing a baseline prosody transfer model to models trained under the two proposed schemes, **text-based** and $\boldsymbol{F_0}$**-based**. **Daft-Exprt** (Zaïdi et al., 2022) was chosen as the baseline since it was a state-of-the-art PT model, representative of the PT field as a whole. Throughout this chapter, **Daft-Exprt** refers to this baseline.

Like the models employing the proposed training schemes, both **Daft-Exprt** and **shuffle** were trained on the Parallel audiobook corpus. Some utterances in the corpus are disproportionately long, which may negatively impact training. So, text prompts longer than 200 characters were removed, leaving 75,267 spoken renditions of 16,275 different texts. Following Zaïdi et al. (2022), 80-bin mel-spectrograms are extracted from recordings and phone durations were estimated using Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). Log-F_0 was estimated using REAPER[1] and l^2-norm of spectrogram frames extracted for energy. Energy and log-F_0 were normalised per speaker. Each model was trained for 24 hours on 16 NVIDIA A100-SXM-80GB GPUs, with a batch size of 192. A pretrained HiFi-GAN (Kong et al., 2020) vocoder was fine-tuned for each trained model.

5.4 Evaluation

5.4.1 Utterance selection

A separate study was conducted to evaluate both hypotheses corresponding to utterance selection, **H5-1a** and **H5-1b**. 55 study participants were asked to indicate which of three references — selected using **text-based**, $\boldsymbol{F_0}$**-based** and **shuffle**— was most prosodically similar to a target utterance. Participants, based in either the US or the UK, were asked to focus on the same prosodic aspects as in the survey employed in Skerry-Ryan et al. (2018). Each participant performed 16 ratings, resulting in 880 responses in total. All participants in this study were recruited and paid via Prolific[2].

[1] https://github.com/google/REAPER
[2] https://www.prolific.co

Different from a typical PT model evaluation, this survey asked participants to compare stimuli that may have different lexical content. To focus participants on only the prosodic content of the stimuli, all samples were delexified using a low-pass filter with a cut-off at 200 Hz and a roll-off of 24 dB per octave. This processing step still allowed participants to judge the qualities used for estimating prosodic similarity in Skerry-Ryan et al. (2018). The survey included two attention checks, where one sample matched the reference while the other two stimuli were noticeably different from the reference. Failing to identify the reference under these conditions would result in poor performance for the rest of the survey. Failing both attention checks resulted in survey elimination. Participant instructions are listed in Appendix B.1.

5.4.2 Model evaluation

Naturalness, preservation of target-speaker identity, and prosody transfer were evaluated for the four models trained under the different training schemes. 60 test text prompts, which had never been seen during training, were used to create the stimuli for the listening tests. Thirty of these prompts were paired with a same-text reference utterance, and the other 30 with a different-text reference utterance. No reference utterance was used as either target or reference during the training of any model. A random target speaker was assigned to each test prompt. Each text prompt was synthesised, using its assigned reference and target speaker, four times, once with each of the four models: **text-based**, **F_0-based**, **Daft-Exprt**, and **shuffle**.

The perceived naturalness of all methods (**H5-2a**) and ground truth utterances was evaluated in a standard Mean Opinion Score (MOS) survey. An equal amount of same-text and different-text synthesised samples was rated for each model. These were created using 30 held-out text-prompts, resulting in 120 evaluations across the four models. 30 ground truth utterances were also included, both the original audios and HiFi-GAN vocoded versions, to gauge possible reduction in perceived naturalness resulting from vocoding. This results in 180 MOS screens in total.

A discriminative AXY test setup was employed to evaluate target-speaker identity preservation (**H5-2b**). Participants indicated whether the synthesised sample (A) sounds more like a ground-truth sample from the target speaker (X) or a ground-truth sample from the reference speaker (Y). A set of different-speaker, different-text samples was extracted from the test set; 30 speaker AXY screens were generated per model, giving 120 in total.

The perceived prosodic similarity to the reference was evaluated for all models to test **H5-3a** and **H5-3c**. Various objective and subjective metrics are proposed in Skerry-Ryan et al. (2018) for evaluating prosody transfer. These include metrics such as Mel-Cepstral Distortion (MCD) (Kubichek, 1993) and voicing decision error (Nakatani et al., 2008), previously discussed in Section 3.2.3.1. Such objective metrics require a gold standard against which to calculate distortion or error, which is only possible for the same-text, same-speaker PT. Since that was not possible here, I followed Zaïdi et al. (2022) and used a MUSHRA-like test for evaluating PT, using **shuffle** as the anchor. To evaluate prosodic similarity, participants were asked to focus on pitch changes, word stress, speaking rate, and pause lengths. Like in Skerry-Ryan et al. (2018), participants were asked to ignore audio quality, as this evaluation should only assess the models' ability to transfer prosody. All same-text and different-text samples from the test set were evaluated, resulting in 60 MUSHRA-like screens. Participants first listened to the reference. Then they rated the prosodic similarity of four samples, each generated by one of the four evaluated models, on a scale from 0 to 100. Participant instructions for the listening test are included in Appendix B.2.

Each MOS, AXY, and MUSHRA-like screen was rated by at least eight different listeners. Again, only native English speakers from the US or the UK were recruited. Each participating listener completed 36 screens in total, pseudo-randomly chosen from each evaluation category (naturalness, speaker identity, prosody transfer).

5.5 Results

A note on statistical test designs

This thesis employs similar statistical tests in several chapters. These designs are motivated and explained in substantial detail in the current chapter to minimise later repetition.

5.5.1 RQ 5-1: Can we automatically find prosodically-informative utterances in a speech corpus?

Results relevant to the first research question are shown in Figure 5.1. The figure shows the prosodic similarity preference rate, as indicated by study participants. The preference of the two proposed methods was compared to the uninformed model, **shuffle**. The study design induced nested random effects, since each rater listened to a random

subset of utterances, and the design included repeated measures per sample. To account for this variation, and to handle binomial outcomes, a hierarchical Bayesian mixed effects model estimated `system` (**text-based**, $\boldsymbol{F_0}$**-based**, or **shuffle**) preference probabilities, accounting for `rater`-level variation and repeated measures (`utteranceID`) with random intercepts. No fixed-effect predictors were included, as the aim was to estimate the overall difference in preference between the `systems`. The Bayesian model estimated the log-odds of participants choosing either **text-based** or $\boldsymbol{F_0}$**-based** over shuffle. Pairwise comparisons were then performed through an Highest Density Intervals (HDIs) analysis of the posterior distributions of each `system` in relation to **shuffle**. Modelling was conducted using `bambi 0.13.0`[3].

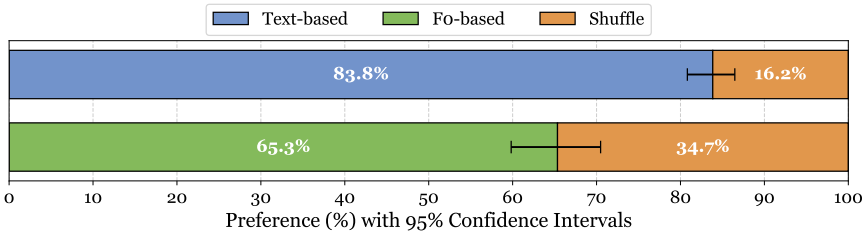


Figure 5.1: Utterances selected by the two proposed methods, **text-based** and $\boldsymbol{F_0}$**-based**, are rated more prosodically similar to the reference than a randomly sampled utterance. This preference is stronger for **text-based** than $\boldsymbol{F_0}$**-based**.

Both proposed methods were associated with an increase in log-odds and, by extension, an increase in the probability of being preferred over **shuffle**. These differences were credible for both $\boldsymbol{F_0}$**-based** (95% HDI: $[0.128, 0.361]$) and **text-based** (95% HDI: $[0.118, 1.432]$) since intervals did not include 0, which would suggest a limited difference between `systems`. In other words, the proposed methods yield training utterances that are informative about the reference prosody:

> **H5-1a**: *Listeners will rate the reference paired by* ***text-based*** *as more prosodically similar to the target utterance than a randomly selected reference.* **Accept** ✔
>
> **H5-1b**: *Listeners will rate the reference paired by* $\boldsymbol{F_0}$***-based*** *as more prosodically similar to the target utterance than a randomly selected reference.* **Accept** ✔

An additional pair-wise comparison between **text-based** and $\boldsymbol{F_0}$**-based** revealed a credible preference for **text-based** (95% HDI: $[0.326, 0.926]$). But both proposed training regimes were evaluated for the downstream PT task.

[3] https://bambinos.github.io/bambi/

5.5.2 RQ 5-2: Does the proposed training regime mitigate the issues of feature entanglement and source-speaker leakage?

5.5.2.1 Perceived naturalness (H5-2a)

Previously, it was predicted that: **H5-2a**: *Speech generated by the proposed methods is perceived as more natural under different-text conditions, when compared to **Daft-Exprt***. The results from the naturalness study are reported in Table 5.1.

Table 5.1: MOS results for recordings, before and after vocoding, and synthesised samples. 95% confidence intervals were estimated using bootstrapping (10 000 resamples).

Model	MOS	.
Ground truth recordings	4.18 ± 0.12	
Ground truth recordings + HiFi-GAN	3.75 ± 0.15	
	Same-text	**Different-text**
shuffle	2.78 ± 0.18	$\mathbf{2.86 \pm 0.21}$
text-based	2.64 ± 0.19	2.40 ± 0.23
F_0-based	2.85 ± 0.22	2.56 ± 0.21
Daft-Exprt	$\mathbf{3.17 \pm 0.20}$	2.42 ± 0.20

To account for `rater` variation and repeated measures, Linear Mixed-Effects (LME) models were employed with random intercepts for both `rater` and `utteranceID`. It should be noted that LME models are typically fitted on continuous data, as opposed to ordinal scale ratings. But, this approximation is frequently made in the TTS literature, for both Multi Stimulus with Hidden Reference and Anchor (MUSHRA) and MOS designs (e.g., Betz et al., 2018; Raitio et al., 2022b; Cohn and Zellou, 2020; Do et al., 2022). LME-based analyses also assume normally distributed residuals and random effects. To validate the design, I confirmed that these were approximately normal and computed the Shapiro-Wilk test statistics (Shapiro and Wilk, 1965). Figure 5.2 shows an example of this for one of the LMEs models employed in the current section. Unless stated otherwise, these approximations were confirmed. All LME-based tests are conducted with `statsmodels=0.14.4` (Seabold and Perktold, 2010).

First, an LME is fitted on MOS ratings of ground truth recordings to account for degradation in quality related to vocoding. `System` (**ground truth** or **vocoded** recordings) is included as a fixed effect with random intercepts for `rater` and `utteranceID`. Assumptions of normality of both residuals and random intercepts were confirmed (Figure 5.2). The estimated mean rating for **ground truth** was 4.21 ± 0.08. The difference

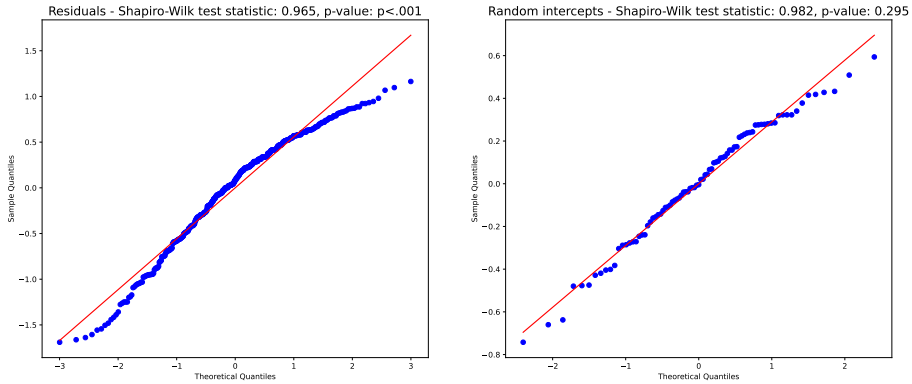


Figure 5.2: QQ-plots for both residuals and random intercepts for the LME model fitted for the **ground truth** vs. **vocoded** comparison. Both were observed to be approximately normal.

between **vocoded** and **ground truth** was estimated as -0.49 ± 0.09, which was significant $(p < .0001, \alpha = .05)$. These less natural, **vocoded** results represent the model's hypothetical best performance.

To evaluate **H5-2a**, a Likelihood Ratio Test (LRT) was conducted. For each pairwise comparison, we fit two LME models: a *full* LME model that includes `system` as a fixed effect — indicating which model generated the utterance — and a reduced model that omits this effect. The LRT compares the goodness-of-fit between the full and reduced model, by evaluating whether including the `system` effect leads to statistical improvements in model fit. The test statistic is derived from the difference in log-likelihoods. Because of this, all LME models are fitted to maximise the full log-likelihood [4]. The test statistic is then compared against a chi-square distribution to obtain a p-value. Random intercepts are, again, included for `rater` and `utteranceID`.

For different text samples, there is no significant difference between **text-based** and **Daft-Exprt** ($\chi^2(1) = 0.00$, $p = .99$). The results are similar for $\boldsymbol{F_0}$**-based** vs. **Daft-Exprt** as no significant difference between the two was observed ($\chi^2(1) = 0.10$, $p = .75$). The hypothesis is rejected as both proposed methods fail to have the predicted improving effect on perceived naturalness:

> **H5-2a**: *Speech generated by the proposed methods is perceived as more natural under different-text conditions, when compared to* ***Daft-Exprt*** — **Reject ✗**

[4] Which is worth mentioning since many software packages, like `statsmodel`, maximise the *restricted* log-likelihood by default.

Compared to **shuffle**, different-text samples from **text-based** were perceived marginally less natural: (-0.423, 95% CI: $[-0.707, -0.139]$), and this difference was significant ($\chi^2(1) = 8.33$, $p = .003$). This was not the case when $\boldsymbol{F_0}$**-based** was compared to **shuffle** ($\chi^2(1) = 2.78$, $p = .09$).

Same-text samples were also evaluated. For these samples, **Daft-Exprt** was rated significantly more natural than both proposed methods: **text-based** ($\chi^2(1) = 14.16$, $p < .001$) and $\boldsymbol{F_0}$**-based** ($\chi^2(1) = 6.91$, $p = .008$). For same-text samples, **text-based** and $\boldsymbol{F_0}$**-based** produced comparably natural-sounding speech ($\chi^2(1) \leq 1.50$, $p \geq .23$). Furthermore, the proposed models are both statistically comparable to **shuffle** in the same-text condition ($\chi^2(1) \leq 1.0$, $p \geq .31$). Taken together, across inference conditions, the proposed models are either perceived as less natural or equally natural as **shuffle**.

Inference condition (same-text vs. different-text) does not significantly affect perceived naturalness for either $\boldsymbol{F_0}$**-based** or **text-based** ($\chi^2(1) \leq 3.30$, $p \geq .07$). However, there is an effect of inference condition for **Daft-Exprt**: (-0.757, 95% CI: $[-1.013, -0.501]$), and this difference is significant ($\chi^2(1) = 31.0$, $p < .001$). In summary, the inference condition has a significant effect on the naturalness of **Daft-Exprt**, but not for the two proposed models.

5.5.2.2 Preservation of speaker identity (H5-2b)

Corresponding to **RQ 5-2**, the proposed training methods were predicted to improve the preservation of the target speaker identity. Figure 5.3 shows listeners' AXY classification accuracy for the four evaluated models. Excluding **Daft-Exprt**, all models have a high speaker classification accuracy, indicating that these models successfully preserve the target speaker identity in different-speaker PT. Participants correctly matched utterances generated by **text-based** to the reference speaker in $196/222$ (88.3%) of cases. This rate is much lower for **Daft-Exprt** where participants only achieved this in $105/228$ (46.1%) of cases. For $\boldsymbol{F_0}$**-based** the classification rate is even higher: $205/224$ (91.5%). The uninformed model achieves the best classification accuracy $205/224$ (91.5%).

The results show a clear difference between **Daft-Exprt** and the other models. But, like other results reported in this thesis, this difference was statistically evaluated.LME models in `statsmodels` do not generally support binary outcomes. So, a Bayesian

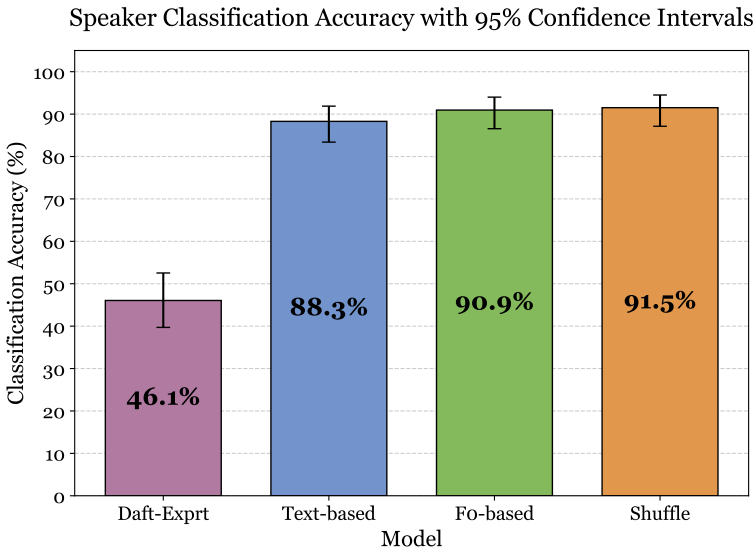


Figure 5.3: The discriminative speaker classification accuracy results, based on subjective AXY ratings, indicate that **Daft-Exprt** fails to preserve the target speaker identity.

mixed effects model was employed instead, using `bambi 0.13.0`. `System` was included as a fixed effect while *rater* and `utteranceID` were random intercepts to account for repeated measures. The Bayesian model estimated a posterior distribution over the log-odds of correct speaker classification (converted to binary outcomes) for each `system`. To assess differences between `systems`, the 95% posterior HDIs were examined. As expected, the differences between **Daft-Exprt** and other `systems` were credible, with the 95% HDIs for all pairwise comparisons excluding 0 (max. HDI range: $[-3.271, -2.165]$). No credible differences were observed between the other three evaluated `systems` (min. HDI range: $[-0.498, 0.371]$).

I therefore confidently accept the hypothesis:

> **H5-2b**: *The proposed methods will better preserve the target speaker identity under different-speaker conditions, when compared to **Daft-Exprt*** **Accept** ✔

However, the results for **Daft-Exprt** were concerning, as they indicated substantial source-speaker leakage for this baseline model. The relative importance of the speaker classification loss can be tuned through the hyperparameter λ_a in Equation 3.2 (p. 48). Increasing the importance of this loss component should discourage source-speaker leakage. Section 5.6.1 provides a discussion on what effects changing this parameter has on the model as a whole.

5.5.3 RQ 5-3: Can a model trained with different yet prosodically similar references retain the level of prosody transfer demonstrated by a baseline prosody transfer model?

MUSHRA-like scores are reported for same-text and different-text prosody transfer in Table 5.2. I predicted that the suggested training regimes would perform PT similarly to **Daft-Exprt** (**H5-3a** and **H5-3b**), while **shuffle** would perform significantly worse than the baseline (**H5-3c**).

Table 5.2: PT MUSHRA-like scores, and 95% confidence intervals, for all models across both inference conditions. The table also includes the previously reported target-speaker classification accuracy.

Model	MUSHRA-like		Speaker classification
	Same-text	Different-text	
shuffle	38.96 ± 5.00	25.45 ± 3.65	91.5%
text-based	38.74 ± 5.18	30.43 ± 3.77	88.3%
F_0-based	42.94 ± 5.30	28.41 ± 3.66	90.9%
Daft-Exprt	61.49 ± 5.68	49.29 ± 4.77	46.1%

Similar to the strategy employed in Section 5.5.2.1, a LME model is fitted on the MUSHRA-like ratings, including random intercepts for both `reference` and `rater`. The LME model is fitted on MUSHRA ratings, where `system` (TTS model + inference condition) is included as a fixed effect, which allowed for intra-model comparisons across inference conditions.

In general, **Daft-Exprt** outperformed all other models in this evaluation. In the same-text condition, there was a significant ($\alpha = .05$) difference between **Daft-Exprt** ratings and all other `systems` ($\chi^2(1) \geq 45.5$, $p \leq .001$). The same was true in the different-text condition, **Daft-Exprt** was again rated better than all other `systems` ($\chi^2(1) \geq 124.1$, $p \leq .001$):

> **H5-3a**: *The prosodies generated by **text-based** and **Daft-Exprt** are equally similar to that of the reference* **Reject ✗**

> **H5-3b**: *The prosodies generated by **F_0-based** and **Daft-Exprt** are equally similar to that of the reference* **Reject ✗**

There was no significant difference between the proposed training regimes and **shuffle** for same-text samples ($\chi^2(1) \leq 1.73$, $p \geq .18$). For different-text samples, however, the

proposed methods were rated marginally better than **shuffle** ($\chi^2(1) \geq 4.44$, $p \leq .003$). Based on these mixed results:

> **H5-3c**: *Compared to **Daft-Exprt**, **text-based**, and **F_0-based**, the prosody generated by **shuffle** is less similar to that of the reference* **Reject ✗**

shuffle was rated worse than all other `systems` in the different-text condition. Across the board, all models performed better for same-text inputs than different-text ones. For **Daft-Exprt**, there was a significant ($\chi^2(1) \geq 14.49$, $p \leq .001$) reduction in perceived prosodic-similarity for different-text inputs (-11.9294, 95% CI: $[-21.051, -6.806]$). This reduction should not be surprising, as the same-text condition matches its training condition.

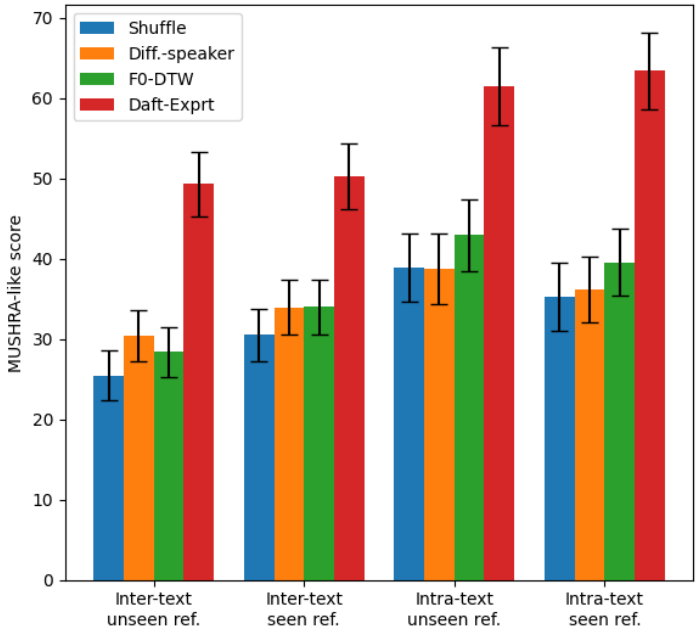


Figure 5.4: MUSHRA-like mean scores for the PT task using both held-out references and ones seen during training. Here, *inter-text* and *intra-text* refer to different-text and same-text, respectively.

Hypothetically, references seen during training should be the easiest use case for the models. So, I hypothesised that if the proposed models had learned transferable representations of prosody under *any* condition, it would be when using known references. However, transfer using seen references does not have a noticeable impact on PT performance, as Figure 5.4 demonstrates. The figure compares the results from Table 5.2 to PT results using seen references. Again, **Daft-Exprt** substantially outperforms the other models using seen references, and the proposed methods are comparable to **shuffle**. This further indicates that the proposed methods have failed to learn to transfer prosody from the reference to the output.

5.5.3.1 F_0-based correspondence between reference and output

text-based and **F_0-based** perform poorly in the PT task compared to **Daft-Exprt**. To better understand why participants preferred **Daft-Exprt**, I objectively analysed the predicted F_0 contours for all four models with two metrics. The first is a normalised frame-level DTW F_0 alignment metric, to evaluate how well the models transfer the reference F_0 contour. The other metric, mean absolute F_0 error, was used to evaluate how well the models preserve the mean F_0 of the target speaker. An evaluation set was created using all corpus speakers and texts from a held-out 60-sentence test set.

Table 5.3: A normalised DTW-based metric indicates how well F_0 contours align with the reference F_0 contour (lower is better, 0 indicating perfect alignment, 1 indicates worst alignment). Mean absolute F0 error indicates how well each model preserves the target-speaker identity.

	F_0 DTW target error		**Mean F_0 target error**	
	same-text	different-text	same-speaker	different-speaker
shuffle	0.60	0.90	19.9 Hz	20.6 Hz
text-based	0.60	0.95	16.9 Hz	18.2 Hz
F_0-based	0.50	0.85	25.7 Hz	20.9 Hz
Daft-Exprt	0.35	0.45	25.4 Hz	43.5 Hz

The results are shown in Table 5.3, broken down by error metric and inference condition. Out of all models, the F_0 contours predicted by **Daft-Exprt** are most similar to the reference, for both same- and different-text samples. This difference is more apparent for different-text samples.

In same-speaker prosody transfer, all models accurately capture the mean F_0 of the target speaker. In the different-speaker case, however, **Daft-Exprt** performs substantially worse than the other models. For **Daft-Exprt**, the mean F_0 of the output is driven towards the mean F_0 of the reference. This may explain the poor speaker preservation results for **Daft-Exprt** presented in Section 5.5.2.2. In summary, **Daft-Exprt** is the only model that closely transfers the reference F_0 contour. But, as evidenced by the different-speaker samples, it fails to preserve the speaker identity while doing so.

5.6 Post-hoc reflections

The current study closely followed (Zaïdi et al., 2022) regarding data pre-processing and choice of hyperparameters for all evaluated models. However, some modelling

decisions and hyperparameter choices may have impacted the results. In a post-hoc exploratory study, I looked closely at three critical aspects of the baseline model, **Daft-Exprt**:

1. the reference embedding capacity
2. the relative importance of the reference-speaker classification sub-loss
3. the features used for conditioning the reference encoder.

PT models are often evaluated only under one inference condition, focusing either just on different-text or just different-speaker transfer. To provide a more holistic view of the baseline model, this post-hoc analysis comprised three different PT conditions: (1) same-text/same-speaker, which matches how PT models are typically trained; (2) same-text/different-speaker; and (3) different-text/same-speaker. The latter two conditions are more representative of how PT models are typically used during inference, and correspond with the two proposed training regimes, **text-based** and $\boldsymbol{F_0}$**-based**. The same list of 60 held-out reference utterances was used in this analysis, paired with either the matching target and speaker or randomly sampled ones.

5.6.1 Speaker classifier importance

Daft-Exprt jointly trains a speaker classifier, and using gradient reversal, this classifier is intended to eliminate reference speaker information from the reference embedding. The whole model is trained to minimise the loss in Equation 3.2 (p. 48), where the term $\mathcal{L}_a$ denotes the cross-entropy loss of the speaker prediction and λ_a is a hyperparameter used to regularise the relative importance of this sub-loss when training the model. Zaïdi et al. (2022) perform a small study to find a value of λ_a that makes classifying the reference speaker from the prosody embedding difficult while maintaining a high degree of perceived prosody transfer. They settled on a value of $\lambda_a = 0.01$, which was also used in the current study.

In the original work, they only evaluate same-text/different-speaker samples to determine a suitable value of λ_a. As part of this post-hoc analysis, I additionally compared how well the generated representations captured the F_0 contour of the reference, to estimate how λ_a would affect the quality of prosody transfer. Five different settings of λ_a were tested. A higher value of λ_a indicates higher relative importance of the speaker-loss and should, therefore, result in less source-speaker leakage. This relationship was confirmed as illustrated in Figure 5.5. The two lowest λ_a values, including the one

suggested in Zaïdi et al. (2022), generate different-speaker samples with a noticeably higher utterance-level mean F_0 error—suggesting substantial source-speaker leakage. This error is comparable to the same-speaker condition when $\lambda_a = 0.1$ or higher, indicating preservation of the target speaker identity.

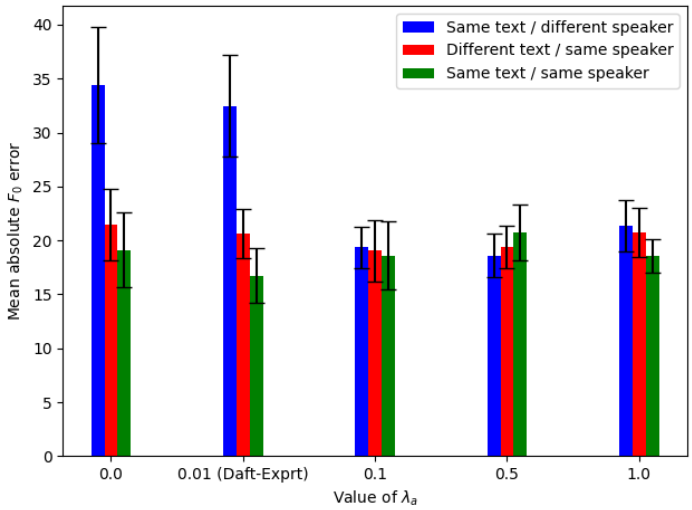


Figure 5.5: Effect of target speaker classification sub-loss on mean F_0. Whiskers indicate the standard deviation across samples within each group.

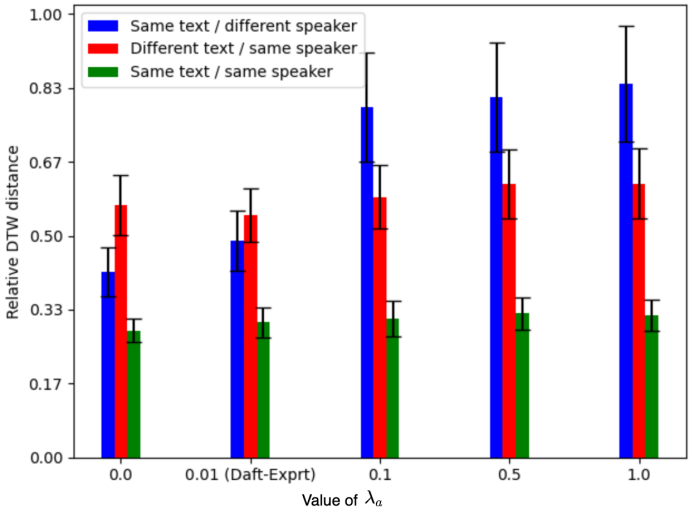


Figure 5.6: Effect of target speaker classification sub-loss on F_0 utterance contour alignment. DTW distances are reported in relative terms.

However, Figure 5.6 shows that the choice of λ_a also affects how well the model captures the shape of the reference F_0 contour. Models trained using λ_a values high enough to preserve the target speaker identity ($\lambda_a \geq 0.1$), generate speech that aligned substantially worse with the reference utterance F_0 contour. Taken together, this suggests that

a higher value of λ_a could have, to a degree, mitigated the **Daft-Exprt**'s observed source-speaker leakage in the main study. However, the trade-off reflected in Figures 5.5-5.6 indicates that such a value would have negatively impacted the quality of prosody transfer.

5.6.2 Transfer capacity

The PT task results, presented in Section 5.5.3, indicated that the proposed training regimes perform similarly to an uninformed model. The performance similarity suggests the baseline PT model fails when trained using non-matching references. In my study, I used the reference embedding dimensionality, $d = 128$, proposed in the original work (Zaïdi et al., 2022). This hyperparameter represents the transfer capacity of the model, and different values could have affected the results presented in the main study. To evaluate this, four **Daft-Exprt** models, where the reference matches the target, were trained using four different embedding capacities. These four models were then evaluated using the two F_0 objective metrics, used in Section 5.5.3, to estimate their transfer capabilities.

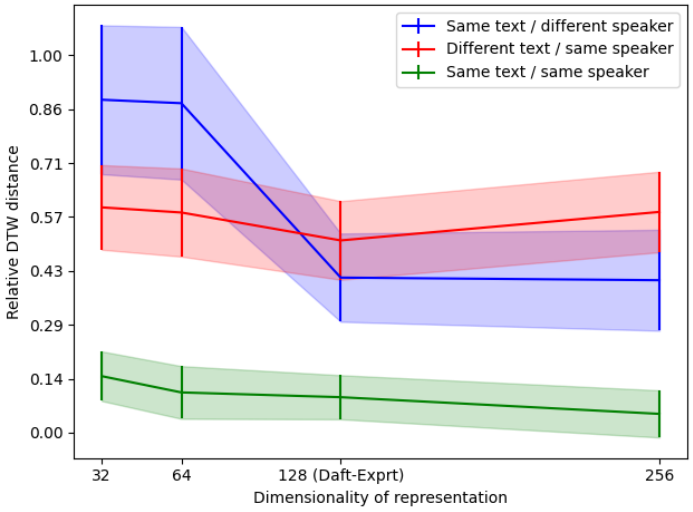


Figure 5.7: Effect of embedding capacity on F_0-contour alignment with the reference (lower is better),

Figure 5.7 shows what effect capacity has on F_0 alignment between the synthesised output and the reference. Capacity seems most salient for cases where the target speaker differs from the reference speaker. A low capacity representation results in different-speaker F_0 contours that align poorly with the reference F_0 contour. This effect is less pronounced for the same-speaker samples. Speaking rate differences could disproportionately penalise different-speaker samples under the DTW-based evaluation.

However, Table 5.4 shows that different-text samples have substantially higher duration differences with the reference. So, duration-related speaker differences do not explain the results in Figure 5.7.

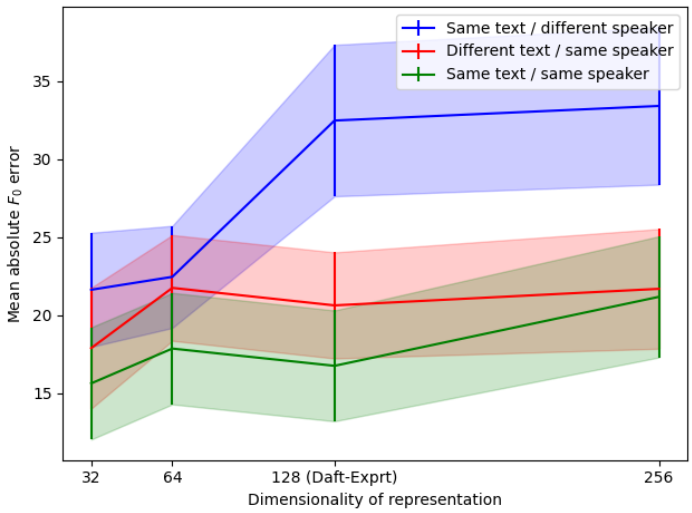


Figure 5.8: Effect of embedding capacity on mean target-speaker F_0 error (lower is better).

For the model trained using the capacity proposed by Zaïdi et al. (2022), and used in the main study, F_0 alignment to reference is comparable for different-text/same-speaker and same-text/different-speaker samples; this confirms that the proposed capacity supports F_0 contour-shape transfer for cross-text input. This also matches with findings from the main study, shown in Table 5.3 (p. 75). Increasing the capacity does not seem to improve F_0 alignment with the reference. Therefore, it is unlikely that increasing or decreasing the representational capacity would benefit models trained using the two training regimes proposed in the main study.

Table 5.4: Mean frame-count difference between synthesis and reference for the three models.

Capacity (d)	**same-text same-speaker**	**same-text different-speaker**	**different-text same-speaker**
32	36.7 ± 17.8	65.4 ± 26.5	301.9 ± 66.2
64	43.6 ± 16.2	64.6 ± 28.9	312.5 ± 74.2
128	47.8 ± 16.8	52.1 ± 23.7	282.7 ± 89.3
256	31.7 ± 15.4	51.6 ± 14.9	309.3 ± 61.1

The main study also showed that **Daft-Exprt** has noticeable source-speaker leakage. Figure 5.8 indicates that increasing modelling capacity results in further source-

speaker leakage. Taken together, the results shown in Figures 5.7 and 5.8 demonstrate a representational capacity trade-off: between either preserving the target speaker identity or enabling the model to transfer prosody.

5.6.3 Choice of reference encoder inputs

As proposed by Zaïdi et al. (2022), the reference utterance mel-spectrogram, frame-level F_0 and energy contours (both speaker-normalised) are all used as inputs for the reference encoder. In Chapter 4, I showed that using the reference mel-spectrogram is not a necessary reference encoder input, and dropping it can be beneficial (Table 4.2, p. 58). To evaluate potential differences based on reference encoder inputs, the **Daft-Exprt** baseline was compared to two models: one that only uses the mel-spectrogram and another that only uses the energy and F_0 contours.

Results based on the same two objective F_0-based metrics are illustrated in Figures 5.9-5.10. Even though the frame-level F_0 sequence is speaker-normalised, the model trained only on reference utterance F_0 and energy contours still struggles to preserve the target speaker identity. This model, which does not use the mel-spectrogram as a reference encoder input, performs similarly to the other two models trained.

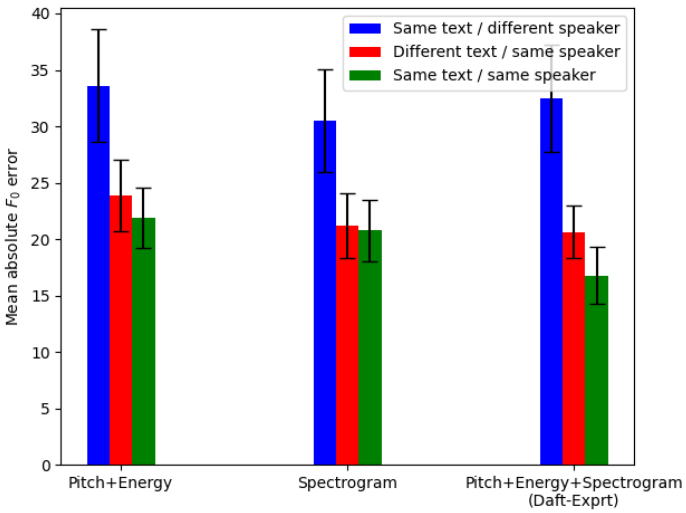


Figure 5.9: Effect of reference encoder input on mean utterance-level F_0 error relative to the target speaker. Lower is better.

Figure 5.10 illustrates how closely the F_0 contours of samples from the three models align with those of the reference utterances. In contrast with the other post-hoc analyses, there is limited difference between the three models in terms of how well they

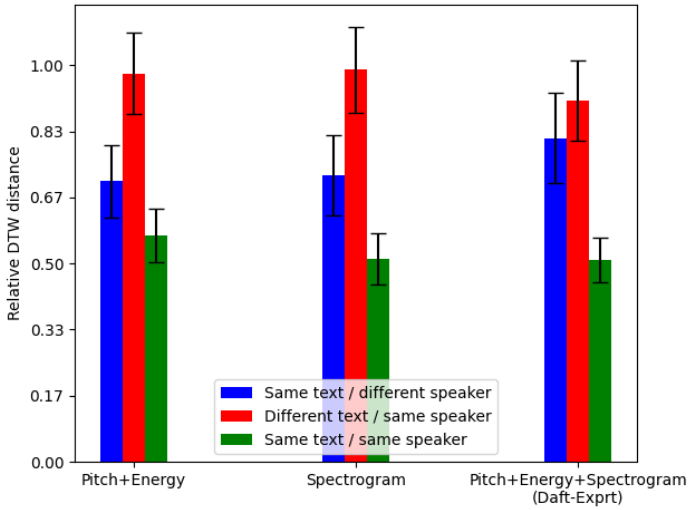


Figure 5.10: Effect of reference encoder input on F_0 utterance contour alignment.

capture the reference F_0 contour, suggesting that the mel-spectrogram is not a necessary input for successful PT model training. But, as evidenced by Figure 5.9, changing the reference encoder inputs would not have mitigated source-speaker leakage.

5.7 Conclusion

In summary, the results demonstrate that:

> **RQ 5-1**: the proposed reference selection methods yield prosodically informative utterances;
>
> **RQ 5-2**: models trained using such references preserve the speaker identity under different-speaker conditions but fail to improve perceived naturalness under different-text conditions;
>
> **RQ 5-3**: the proposed models fail to transfer prosody to the same degree as the baseline model.

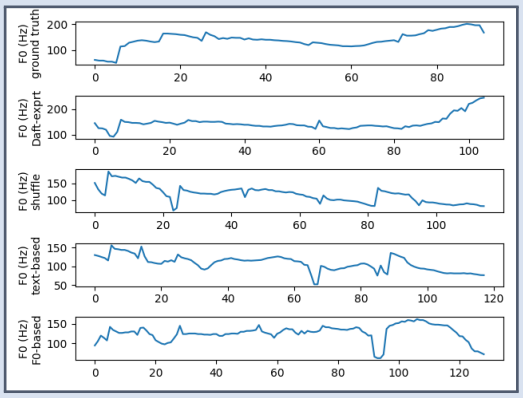
To perform different-text and different-speaker PT, prosody has to be separated from other information in the reference. **Daft-Exprt** aims to address this using modelling techniques to limit or boost the model's access to certain reference features. Still, Sections 5.5.2.1-5.5.2.2 showed that this state-of-the-art PT model suffers from source-speaker leakage in the different-speaker condition, and significantly lower perceived naturalness in the different-text condition. Results from the post-hoc analysis indicate that addressing these issues through modelling decisions, such as embedding capacity or speaker loss importance, is not a sufficient solution.

To address these issues, I instead proposed training the model on different, but prosodically similar references. If PT models, like **Daft-Exprt**, actually modelled transferable representations of prosody, it should be possible to train them on such references. Yet, the results presented in this chapter indicate that this is not the case. Instead, the proposed training methods, **text-based** and **F_0-based**, have a detrimental effect on the quality of perceived prosody transfer.

5.7.1 Transferability of representations

Different-text PT samples generated by **Daft-Exprt** closely imitate the reference prosody at the cost of perceived naturalness. To achieve this, **Daft-Exprt** models a representation of prosody that is highly source-text-dependent, regardless of the target text. Ultimately, this results in decreased perceived naturalness.

The Results of Forced Prosodic Similarity



The figure to the side provides an anecdotal example of the trade-off between PT and perceived naturalness. The frame-level F_0 contour of a ground truth utterance (*"Did you become engaged then?"*) is shown in the first sub-figure. This reference was spoken using the typical rising intonation of a question. I then synthesised samples for all evaluated models while conditioning on a reference that was not a question (*"The top of the morning to you Jim."*). The corresponding F_0 contours are shown in the same figure. Only **Daft-Exprt** transfers the rising intonation, which sounds natural for the reference question but highly unnatural for the target text.

Furthermore, as has been demonstrated both subjectively and objectively, **Daft-Exprt** fails to disentangle speaker identity from prosody in different-speaker PT cases. Still, **Daft-Exprt** is rated highest in same-text, different-speaker PT while a model trained under these conditions, **text-based**, performs similarly to an uninformed model (**shuffle**). These results suggest that the prosodic representation modelled by **Daft-Exprt** is highly reference-speaker dependent. Results from the post-hoc analysis—presented in Section 5.6.1—indicate that a lower relative speaker classifier importance (λ_a) could have improved the perceived PT performance of the **text-based** model. However, the same

results showed that a lower λ_a would likely lead to increased source-speaker leakage.

Based on this, one must consider whether the latent spaces constructed by state-of-the-art PT models, like **Daft-Exprt**, are appropriate for transferring prosody. PT approaches rest on the premise that they model a latent space that captures *transferable* prosodic variation from the training data. This is also the key premise used to motivate the proposed training regimes in the study. I believe that the negative results demonstrate that this premise does not hold.

Chapter 6

The role of style in speaker identity judgements

This chapter discusses an alternative application of reference-based control: *voice cloning*. Here, the aim is to capture an unseen speaker's identity from a single reference utterance, which then conditions the entire speech generation process. The work in this chapter is based on the co-authored paper *Just Because We Camp, Doesn't Mean We Should: The Ethics of Modelling Queer Voices* (Sigurgeirsson and Ungless, 2024). In this paper, we evaluated a *voice cloning* model's ability to capture a speaking style colloquially known as "gay voice". We approached this work as authors who both identify as gay men. As reflected in the title, the paper explores the ethical implications of voice cloning models, particularly when used to replicate queer voices. My co-author, Eddie Ungless, contributed most of this ethical discussion. Therefore, the scope of the current chapter is limited to the voice cloning evaluation performed in the paper, which I devised. The objectives of voice cloning differ from those of Prosody Transfer (PT). For example, voice-cloning models do not explicitly aim to transfer prosody from the reference. But, our results show that the perception of a particular speaking style — expressed using a mixture of phonetic and prosodic traits — informs the perception of speaker identity.

6.1 Research objective

Many people hold stereotypes about what gay men sound like: "soft", with a "lisp", feminine (see Mack and Munson (2012)). Without affirming these judgements, we

acknowledge that there is evidence of phonetic features that differ between gay and non-gay speakers, some of which influence perceived speaker sexuality. No single set of features is always present in *gay voice* (Mack and Munson, 2012; Smyth et al., 2003), but rather gay male speakers are more likely to exhibit some particular phonetic practices, e.g. hyper- (Munson et al., 2006) and mis-articulated /s/ (Mack and Munson, 2012); longer sibilant duration and greater pitch range also impact perceived gay sexuality (Levon, 2007). In this study, we asked:

> **RQ 6-1:** *Can perceived gay voice be successfully modelled by an end-to-end voice-cloning TTS model?*

Any speaker identity modelling necessarily involves some reduction, and we believe this will be exacerbated in the case of men with *gay voice*. Existing speech datasets are not focused on capturing queer voices; thus men with *gay voice* will likely be vastly outnumbered by those who do not (*control* speakers), if represented at all. Therefore, we believed that because these models implicitly do not model *gay voice*, the resulting output would be poor in terms of fidelity to source voice. As such, we predicted:

> **H6-1a:** *Compared to the difference between ground truth and synthesised control speaker utterances, synthesised gay voice utterances will show a larger decrease in perceived gayness relative to their ground truth utterances.*

Because we believe perceived gayness is an important part of speaker identity, we also believe that this loss in perceived gayness will be associated with a loss in perceived speaker similarity. As such, we predict:

> **H6-1b:** *Compared to the difference between ground truth and synthesised control speaker utterances, synthesised gay voice utterances will be rated as having a lower speaker similarity with ground truth utterances.*

In this work, we also explored answers to:

> **RQ 6-2:** *Which steps in a speech synthesis pipeline have the biggest impact on the fidelity of gay voice?*

6.2 Stimuli creation

There are no available Text-To-Speech (TTS) corpora labelled by sexuality that we were aware of. We resorted to selecting speakers with *gay voice* from a large generic speech corpus, Ted-Lium 3 (Hernandez et al., 2018), which includes speech from several self-identifying gay men. All were speakers of US American English. This narrow scope ensured familiarity for us as data curators and our annotators. We manually selected speakers based on our shared perception of them having *gay voice* (henceforth, `gay` group). Speakers were included if both authors agreed the speaker had *gay voice*, although the perceived "strength" of *gay voice* varies across selected speakers. We confirmed that all selected speakers publicly identify as gay and/or are in public relationships with men.

For our control data, we chose speakers who we perceived not to have *gay voice* (henceforth, `control` group) and who do not publicly identify as gay, acknowledging that these speakers may nonetheless be gay: their sexuality is not directly relevant to whether *gay voice* is captured by speech synthesis models. We did not control for regional accent variation, talk venue (which impacts quality) or talk topic, which may influence *gay voice* perception (Kachel et al., 2023). Our data consisted of speakers aged 29-58 at the time of recording. We included 20 white and 4 Black speakers (perceived race). We chose approximate controls for each *gay voice* speaker: similar age and recording year; we paired Black `control` speakers with Black `gay` speakers. To the best of our knowledge, all speakers are men.

We studied the perception of *gay voice* at different stages of TTS development. We considered two different ground-truth conditions (henceforth, clip types) for evaluation. First, long utterances: we selected three 30 s segments (henceforth **Segment**) per speaker, ensuring none contained discussion of LGBT+ topics. TTS models are typically not trained on such long utterances, but we anticipated that longer segments would better capture speaker identity. We then considered ground-truth segmented utterances (henceforth **Raw-utt**), as segmented in TED-Lium 3, which reflect typical TTS training data. We selected 15 such utterances per speaker, excluding any which contained terms from a list of queer vocabulary (comprised of a list taken from Wikipedia[1] and supplemented through our own knowledge) to minimise the likelihood of sexuality disclosure.

[1] https://en.wikipedia.org/wiki/LGBT_slang

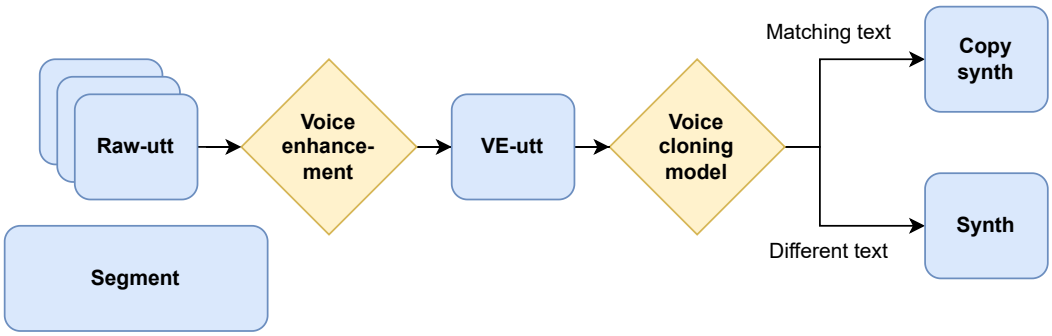


Figure 6.1: Flowchart illustrating the TTS pipeline we employed and the different stages (blue) we evaluated. We evaluated both long **Segments** and short **Raw-utt** ground truth utterances. **Raw-utt** are passed through a voice enhancement model to produce **VE-utt**. Using a voice-cloning model, we synthesise utterances with either a text that matches the one in the reference (**Copy-synth**) or not (**Synth**).

The Ted-Lium 3 corpus is not purposefully made for TTS modelling, and many utterances include background noise, music and unlabelled speech. For our third clip type, we employed the SepFormer voice-enhancement model (Subakan et al., 2021) trained on the Microsoft-DNS 4 voice-enhancement challenge corpus (Dubey et al., 2022). This pretrained model was accessed through the *SpeechBrain* toolkit (Ravanelli et al., 2021). We passed all **Raw-utt** through the voice enhancement model to form our third clip type, **VE-utt**.

We finally considered the perceived *gay voice* of synthesised speech. The data we collected was insufficient to train a TTS model from scratch. Instead, we used a pretrained version of XTTS-v2 from Coqui (Eren and Team, 2021). XTTS-v2 is a multi-lingual, multi-speaker TTS model. The model is trained on more than 16,000 hours of publicly available speech data and supports zero-shot reference-based voice cloning. We evaluated two different clip types of synthesised speech: (1) **Copy-synth**, where the target text matched the text in the reference (**VE-utt**) used to generate the zero-shot speaker embedding; and (2) **Synth**, where the target text does not match. We anticipated that **Copy-synth** would represent the optimal conditions for models such as XTTS-v2 to model a speaker's identity.

6.3 Evaluation

6.3.1 Participants

All our participants were L1 speakers of English from the US and the UK. We only recruited LGBT+ participants because we are principally interested in whether synthesised *gay voice* is perceptible to other community members, not whether it is detectable by out-group members, and listener sexuality may influence perceived sexuality (Smyth et al., 2003; Munson, 2011)

6.3.2 Metrics

Metric design for perceived *gay voice* differs across the literature in ways that are likely to be significant (Munson, 2011). We chose our specifications as follows: as we were interested in the perception of *gay voice*, not perceived speaker sexuality, we asked participants whether the voice *sounds* gay rather than whether the speaker *is* gay, similar to Mack and Munson (2012) but counter to e.g. Kachel et al. (2023). We employed a 7-point Likert scale (in line with e.g., Gaudio, 1994; Levon, 2007; Kachel et al., 2023) where 1 is "definitely sounds straight" and 7 is "definitely sounds gay". We also asked participants to rate perceived naturalness using a standard 5-point Likert scale Mean Opinion Score (MOS) format. We employed a *semi*-continuous scale for rating speaker similarity as we suspected that differences between speakers might be more subtle than differences in perceived *gay voice*. We used a scale of 0-100 (0 "completely different", 100 "exactly the same"), between a reference **Segment** and **Synth**.

6.4 Results

6.4.1 Pre-study

We conducted a pre-study to confirm that our `gay` and `control` speakers showed divergent *gay voice* ratings. We recruited 13 participants to rate perceived *gay voice* for 30 **Segments** each, such that each **Segment** was rated approximately 5 times. Based on participant ratings, we discarded 5 speakers from each group, leaving those with a rating above 5 (for `gay`) or below 3 (for `control`).

6.4.2 RQ 6-1: Can perceived gay voice be successfully modelled by an end-to-end voice-cloning TTS model?

6.4.2.1 Preservation of *gay voice* (H6-1a)

To test whether *gay voice* is lost in speech synthesis (**H6-1a**) and at which stage in the pipeline *gay voice* is lost (**RQ 6-2**), each participant rated (a) two 30 s **Segments** and (b) a random subset of 30 utterances (from **Raw-utt**, **VE-utt**, and **Copy-synth**) for how gay the voice sounded. In the same listening test, we evaluated perceived naturalness of (c) **Segments** and (d) utterances (**Raw-utt**, **VE-utt** and **Copy-synth**). Naturalness results are employed in Section 6.4.3 for answering (**RQ 6-2**). All clips were allocated to participants at random, and the order of tasks (a-d) was pseudo-randomised. Participant instructions are provided in Appendix C.1.1. Two participants were discarded for failing attention checks. Each utterance was rated approximately 4 times.

Table 6.1: Mean rating $\pm$ standard deviation for naturalness and for gay voice, by condition and by group.

	Naturalness		Gay Voice	
Group	`Gay`	`Control`	`Gay`	`Control`
Segment	3.6 ± 1.2	4.2 ± 0.8	5.3 ± 1.3	2.6 ± 1.3
Raw-utt	3.5 ± 1.2	3.9 ± 1.0	5.7 ± 1.3	2.7 ± 1.5
VE-utt	3.4 ± 1.2	4.0 ± 1.0	5.7 ± 1.3	2.5 ± 1.3
Copy-Synth	2.7 ± 1.3	2.9 ± 1.3	4.7 ± 1.5	3.3 ± 1.6

Results are given in Table 6.1, broken down by speaker in Figure 6.2, and by clip type in Figure 6.4 (p. 93). We used `lme4` (Bates et al., 2015) to fit Linear Mixed-Effects (LME) models, and `afex` (Singmann et al., 2024) to conduct likelihood ratio testing for p-values — appropriate given nested random variables with many levels. As is to be expected, `control` speakers were rated as having much lower *gay voice* than `gay` speakers: the estimated effect was -2.59 ± 0.26 according to a LME model with group as fixed effect, and annotator, speaker and utterance ID as random intercepts. This was significant ($\chi^2(1) = 26.02$, $p < .001$).

Comparing mean ratings for **Segments** versus **Copy-synth** for `gay` speakers, we found that **Copy-synth** clips were rated to sound much less gay. Unexpectedly, ratings *increased* for `control` speakers. We speculate that this may be because, for straight men, read speech of the kind typically used to train TTS models is rated as more gay-

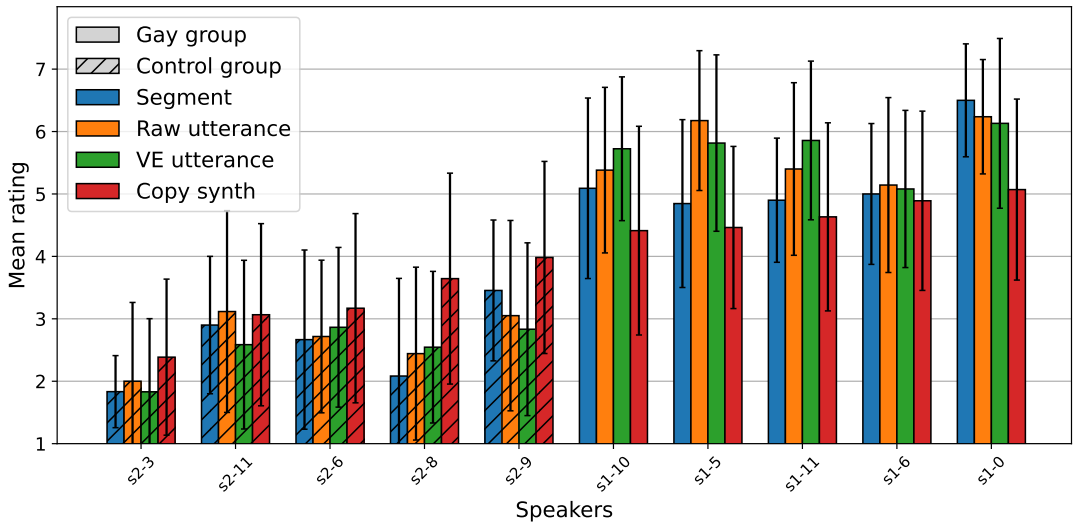


Figure 6.2: Mean *gay voice* ratings for all evaluated speakers. Here, *Set 1* refers to `gay` speakers and *Set 2* refers to `control` speakers.

sounding than spontaneous speech (Smyth et al., 2003).

An LME model fitted on *gay voice* ratings for **Segments** and **Copy-synth**, with group and clip type as fixed effects, and listener, speaker and utterance ID as random intercepts, found a main effect of group, estimated as -2.68 ± 0.38 ($\chi^2(1) = 18.52, p < .001$), a non-significant effect of clip type and a significant interaction (1.20 ± 0.31, $\chi^2(1) = 14.26, p < .001$). In a sense, our prediction that `gay` speakers would show a larger decrease in ratings is supported, in that the ratings decreased relative to ratings of `control` speakers:

> **H6-1a**: *Compared to the difference between ground truth and synthesised control speaker utterances, synthesised gay voice utterances will show a larger decrease in perceived gayness relative to their ground truth utterances.* **Accept** ✔

6.4.2.2 Preservation of speaker identity (H6-1b)

To test whether loss in perceived gayness is associated with a loss in speaker similarity (**H6-1b**) we presented participants with a 30-second **Segment** for a speaker as reference, plus a corresponding **Synth** for that speaker alongside two anchor utterance from other speakers (one each from `gay` and `control`), and asked them to rate the utterances on similarity. Similarity accuracy was measured as the frequency with which the correct sample was selected as the most similar. **Segments** and **Synth** were separately

rated on naturalness and *gay voice*. Each participant rated 20 speakers' utterances; each utterance was rated 4 times. Clips were randomly allocated to participants. Task order was randomised. Participant instructions are listed in Appendix C.1.2. One participant was discarded for failing attention checks.

We conducted a Kruskal-Wallis test (McKight and Najab, 2010), as our residuals were not normally distributed, and found that there is no significant difference in similarity accuracy between the groups, so **H6-1b** is not supported:

> **H6-1b**: *Compared to the difference between ground truth and synthesised control speaker utterances, synthesised gay voice utterances will be rated as having a lower speaker similarity with ground truth utterances.* **Reject ✗**

A Spearman rank-order correlation (Dodge, 2008, p. 502-505) was run to estimate the effect that loss of speaking style has on the loss of speaker identity. When we examined the correlation between *gay voice* rating and similarity, we noticed that for `gay` speakers there was a medium correlation between size of loss of *gay voice* and similarity accuracy ($r_s = -3.44, p = .021$), and no such correlation for `control` speakers. It seems loss of perceived *gay voice* for `gay` speakers is more pertinent to similarity judgements than "loss" of "straight voice".

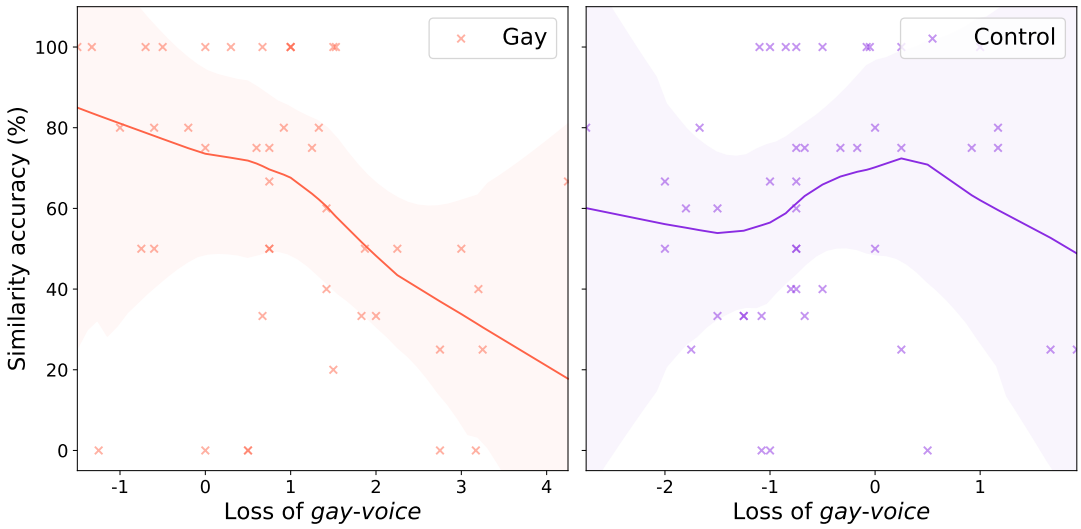


Figure 6.3: Graphs showing the correlation between the similarity accuracy and reduction in *gay voice* between reference **Segments** and **Synth** (A positive value indicates a loss). Lines represent LOWESS-smoothing (Cleveland, 1981) and the shaded area indicates the 95% confidence intervals.

6.4.3 RQ 6-2: Which steps in a speech synthesis pipeline have the biggest impact on the fidelity of gay voice?

Looking at average *gay voice* rating by clip type to explore **RQ 6-2**, we observed that ratings actually increase between **Segment** and **Raw-utt**, particularly for `gay` speakers. The reduction may have been due to sexuality disclosure, despite our attempts to prevent this. There was not a substantial difference between **Raw-utt** and **VE-utt**, for either group. There is a noticeable drop in ratings between **VE-utt** and **Copy-synth** for `gay` speakers and an increase for control speakers.

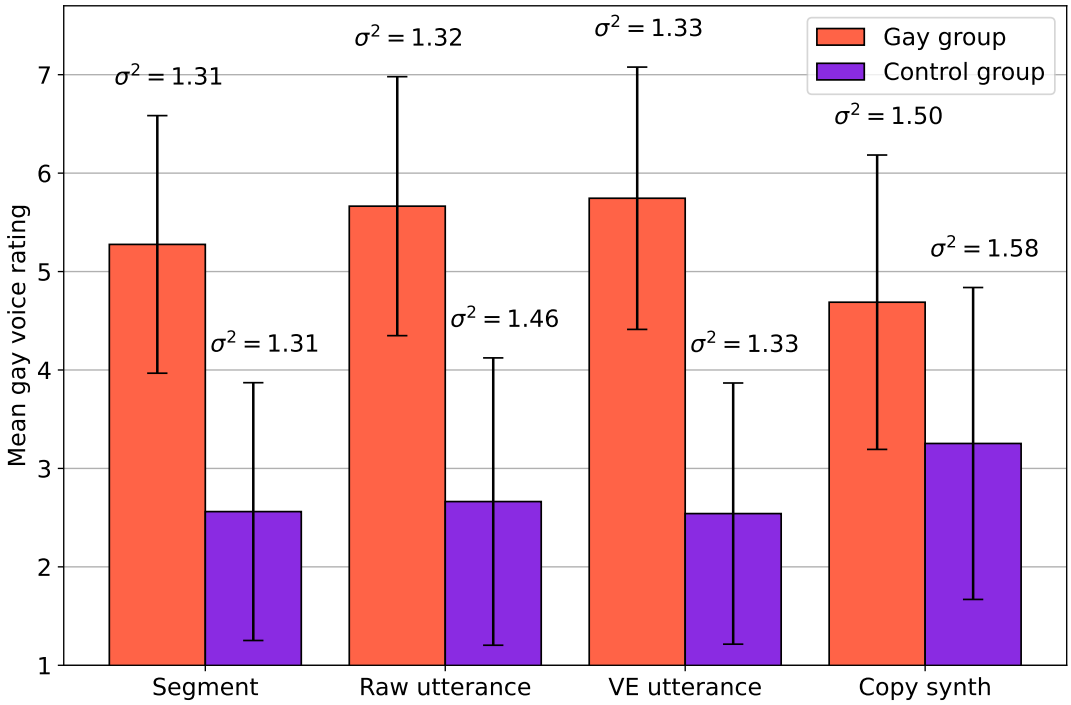


Figure 6.4: Mean gay voice rating of both speaker groups in each clip type.

We fitted a series of pairwise LME models with clip type and group (`gay` or `control`) as fixed effects and listener, speaker and utterance ID as random effects. We then tested for significant differences in mean *gay voice* ratings between *clip types* at consecutive steps in the TTS pipeline: **Segment** → **Raw-utt** → **VE-utt** → **Copy-synth**. We evaluate both ground-truth clip types (**Segment** and **Raw-utt**) since we anticipated that longer segments would better reflect the use of *gay voice*.

Between **Segment** and **Raw-utt**, clip type had a weak significant effect estimated as 0.44 ± 0.20 ($\chi^2(1) = 3.91, p = .048$), main effect of group was -2.71 ± 0.38 ($\chi^2(1) = 23.65, p < .001$); a non-significant interaction between group and clip type. Between

Raw-utt and **VE-utt**, there was a non-significant impact of clip type, main effect of group was -3.03 ± 0.25 ($\chi^2(1) = 28.38, p < .001$); also a non-significant interaction. Between **VE-utt** and **Copy-synth**, there was a significant effect of clip type, estimated as -1.49 ± 0.23 ($\chi^2(1) = 4.50, p = .034$); a main effect of group at -1.49 ± 0.23 ($\chi^2(1) = 25.38, p < .001$); and a significant interaction effect at 1.75 ± 0.16 ($\chi^2(1) = 115.38, p < .001$).

In answer to **RQ 6-2**, synthesis had the biggest impact on *gay voice* ratings. Contradicting our prior belief, gay voice ratings slightly increased between **segment** and **VE utterance**. Table 6.1 (p. 90) shows the naturalness judgements of both groups at the different stages of the TTS pipeline. *Gay voice* speakers were rated as less natural on average, which we speculate is due to greater background noise in the data employed for the `gay` group. As is to be expected, perceived naturalness decreases through the pipeline, the largest drop coming at synthesis. But this is true for both groups, which receive similar naturalness judgments at the synthesis stage. It is, therefore, unlikely that degradation in naturalness caused any differences in *gay voice* ratings between the two groups.

6.5 Conclusion

Modern voice cloning models claim to be able to capture a diverse range of voices. Here, we tested the ability of a typical pipeline to capture a speaking style known as *gay voice*. The results demonstrated a homogenisation effect: synthesised speech is rated as sounding significantly "less gay" than its corresponding ground-truth for speakers with *gay voice*, but ratings actually increased for control speakers.

We also found that for speakers with *gay voice*, loss in this speaking style corresponded to, on average, lower speaker similarity ratings. At the same time, we observed no such correlation for the `control` speakers. This is an important point: it seems that for some speakers, the speaking style they employ forms a salient aspect of their speaker identity. How people use prosody, which may reflect a choice of speaking style, plays a vital role in the perception of speaker identity (Helander and Nurminen, 2007). Correspondingly, failing to model speaking styles is a failure in modelling their identity. As this is the case for the model evaluated here, this points towards a limitation in typical zero-shot voice cloning models.

Chapter 7

Discussion

This part investigated reference-conditioning as a way to control Text-To-Speech (TTS) models. Reference-conditioning can be used to guide the prosodic generation for a broad spectrum of TTS tasks, as evidenced by the prior literature (e.g., Skerry-Ryan et al., 2018; Oplustil-Gallegos and King, 2020; van Rijn et al., 2021; Valle et al., 2020a). Reference-based control may be considered intuitive, as the control signal (the reference) matches the modality of the output. Therefore, reference-based control enables the communication of control intent without requiring any intermediate steps. However, finding a suitable reference is non-trivial.

I have primarily focused on applying this control method to the prosody transfer task, which typically aims to provide control over features relevant to linguistic prosody (Section 2.2.1). Prosody is, naturally, both text- and speaker-dependent. Therefore, a transferable representation of prosody must be invariant to the reference speaker and text, such that it can be applied to any target text and speaker. We confirmed in Section 5.5.2.2 that utterance pairs selected by both **text-based** and $\boldsymbol{F_0}$**-based** are prosodically similar. However, this similarity does not appear to be sufficient for the models to transfer prosody. Therefore, Prosody Transfer (PT) models only appear to work if the reference is identical to the target during training. Since the representations predicted by PT models are dependent on both the reference speaker and reference text, I conclude that representations of prosody—predicted in such a way—are not transferable.

As demonstrated in Chapter 6, the feature entanglement issue is not limited to PT. Voice-cloning, a form of reference-based control, aims to control speaker identity

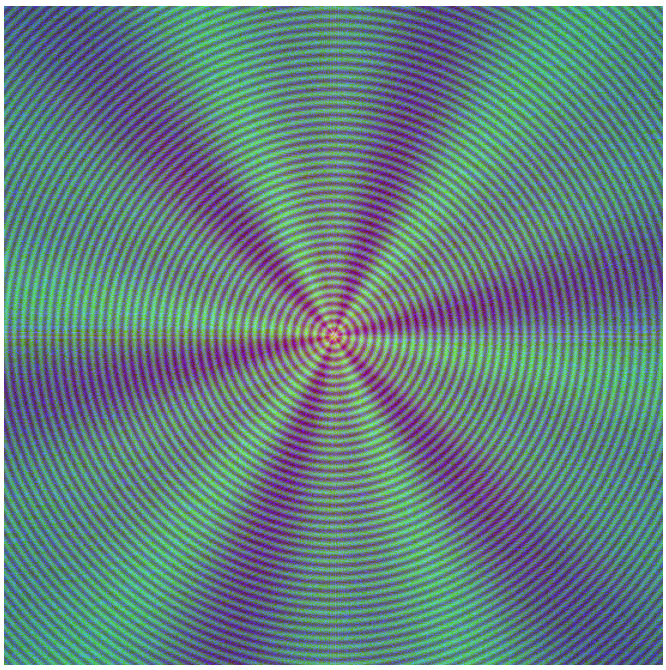
based on limited reference speaker data. Chapter 6 shows that a popular voice-cloning architecture fails to separate speaking style from speaker identity for certain speakers. That, in turn, results in the model's failure to preserve the identity of those speakers when compared to speakers where this feature entanglement is not observed.

PT models aim to provide the temporally-precise control required to dictate prosody generation differently across the utterance. So, PT models typically employ a strategy to transfer temporally fine-grained information from the reference to the output. But as a result, perceived PT naturalness is degraded in the different-text condition since a PT model generates a reference-text dependent representation of prosody. Reference-based control may, therefore, be better suited for tasks where temporally-precise control is not required. For example, Global Style Token (GST)-based models and voice-cloning models, which do not aim to control prosody specifically, learn more generalisable representations than a typical PT model.

One might ask whether tasks like different-text prosody transfer can ever be achieved: can prosody from a different text ever be transferred to another without negatively impacting naturalness? This question is investigated in Chapter 10. The results suggest that human annotators can enhance the perceived quality of different-text PT utterances, indicating an improvement that reference encoders are unable to provide.

Part II

Acoustic Feature Control



Typical end-to-end TTS models are trained to predict a mel-spectrogram directly from text. However, some models take a different approach by separately predicting variations in selected acoustic features. While such feature separation is often motivated by modelling advantages, it also enables interpretable control over specific aspects of speech as a by-product. This part introduces control methods that leverage this approach. I refer to models that support such modifications as Acoustic Feature Control (AFC) models.

Published papers presented in this part:

Chapter 9: *Controllable Speaking Styles Using a Large Language Model* (Sigurgeirsson and King, 2024)

Chapter 10: *A Human-in-the-Loop Approach to Improving Cross-Text Prosody Transfer* (Maurya and Sigurgeirsson, 2024)

Chapter 8

Introduction and background

8.1 Introduction

Part I introduced the concept of *reference-conditioning* as a means to control Text-To-Speech (TTS) models, with a particular focus on its application to Prosody Transfer (PT). In PT, the aim is to guide the output prosody via a reference utterance submitted to the model. Such control might be considered intuitive, as the prosodic specification is provided in the same modality as the output: speech. PT models aim to holistically transfer the reference prosody to the target text. This characteristic reflects on how PT models are trained: the model learns to use *all* information from the reference that it finds helpful in reconstructing the entire target utterance. Previously, Chapter 5 showed that this training regime leads to feature entanglement as PT models fail to model transferable representations of prosody. Beyond this entanglement, the holistic design of PT also prevents temporally-precise control, like the control of individual words or syllables, because the reference conditions the generation of the entire utterance. Further to that, PT models learn combined representations of multiple acoustic features in an inseparable manner. Therefore, PT models cannot, for example, change just the intonation or only the speaking rate. While reference control of prosody can be considered intuitive, it is limited to utterance-level guidance, whereas many practical applications may call for more nuanced control.

Certain TTS models, like **FastSpeech 2** (Ren et al., 2020), separately predict acoustic correlates of prosody. Learning how text inputs correspond to variation in these features alleviates overfitting to particular prosodic renditions found in the training cor-

pus, which improves generalisation (Ren et al., 2020). As a by-product of this training regime, Acoustic Feature Control (AFC) models like **FastSpeech 2** are good candidates for interpretable and precise speech control. AFC control is realised through the modulation of features already predicted by the model. This control is optional, as AFC models can generate speech without any control inputs. In contrast, typical reference-based (Part I) or prompt-based (Part III) methods *always* require some input, in addition to text, to generate speech. **FastSpeech 2** was not specifically proposed as a *controllable* model, and much follow-up work has extended the architecture to enable different types of control (Section 8.2.2). However, models like **FastSpeech 2** can already be controlled, to some extent, without modifications. Hypothetically, a Human-in-the-Loop (HitL) could make adjustments to, say, just the predicted F_0 contour of a particular word by modifying the initially-predicted F_0 values. Such control over predicted features forms the central theme of the current Part.

Chapter 9 presents the results from *Controllable Speaking Styles Using a Large Language Model* (Sigurgeirsson and King, 2024). The study evaluated a *control scheme* that interacts directly with variance predictors, similar to those employed in **FastSpeech 2** (Section 2.3.2.3). The proposed control scheme was tested on several tasks to gauge the scheme's ability to augment an originally predicted prosodic rendition into a more appropriate one for the given target text and optional contextual cues. Instead of employing a HitL to make the appropriate adjustments, an Large Language Model (LLM) is prompted to suggest the adjustments. So, although modifications were made in terms of known acoustic features, a user of the proposed model specifies the desired behaviour in a text-based natural language description instead.

Section 10 covers the work from *A Human-in-the-Loop Approach to Improving Cross-Text Prosody Transfer* (Maurya and Sigurgeirsson, 2024). In this study, we evaluated whether a control scheme, similar to the one presented in Section 9, can be used to improve the quality of *cross-text* PT samples. We tasked HitL participants to make adjustments to F_0, duration, and energy — using the proposed control scheme — to make the output rendition more appropriate for the target text, while keeping it prosodically similar to a given reference. We also assessed the effort involved in completing the task, participants' self-reported success rates, and their level of agreement, in order to evaluate the usability of the proposed control scheme.

Both studies presented in this part discuss other means of TTS control, but through an analysis based on AFC. The first study proposes an approach which is comparable

to some prompt-based TTS models, which are later discussed in Part III. The second study aims to improve the perceived quality of a PT model, extensively discussed in Part I. The difference in these two studies highlights a key advantage of AFC-models: control is achieved through interpretable features instead of learned, latent representations, which may not lend themselves well to interpretation. As demonstrated in the current chapter, this interpretability not only facilitates more transparent control but also enables the analysis of alternative control methods.

8.2 Background

Active human guidance, provided by a Human-in-the-Loop (HitL), has been applied to a wide range of Machine Learning (ML) tasks (Wu et al., 2022). HitL feedback is employed to make use of the broad world knowledge of humans to improve the perceptual performance of ML models (Mosqueira-Rey et al., 2023). This use of HitL guidance is prominent in Natural Language Processing (NLP) tasks, including text classification (Arous et al., 2021), text summarisation (Stiennon et al., 2020) and sentiment analysis (Liu et al., 2021), for example. Often, HitL refers to the incorporation of this world knowledge in the training of the model, through approaches like Active Learning (AL) (Settles, 1995) or Interactive Machine Learning (IML) (Amershi et al., 2014). But it may also refer to an inference process being actively guided by a HitL without any changes to the ML model's learning process (Wu et al., 2022).

The main studies presented in Chapters 9 and 10 align with this second definition, as they explore control methods that incorporate HitL-based guidance during inference, without modifying the model's training procedure. But here I provide a summary of the broader range of HitL applications for TTS, which informed the research presented in this part of the thesis. HitL-based strategies have been employed for several tasks in TTS, such as emotive speech modelling (van Rijn et al., 2021), speaker adaptation (Udagawa et al., 2022), speaker identity modelling (Saito et al., 2021) and modelling speaking styles (Cornille et al., 2022). How HitL participants interact with the TTS model, and which aspects of the inference process are influenced by them, differ between approaches. Here I focus on two main branches of work in TTS that support HitL guidance: (1) Human-based refinement of the learning process through strategies such as AL or feature exploration; and (2) TTS models that are designed explicitly for human-guidance of the inference process.

A note on human-in-the-loop notation

Personally, I find it a little bit awkward to refer to participants as HitLs (Humans-in-the-loop?). I prefer to reserve the term HitL to describe the nature of the task itself. Throughout this chapter, I often use the term HitL *participant* or simply *participant*.

8.2.1 Model refinement through perceptual feedback

8.2.1.1 Augmenting training through human perception

Certain HitL processes aim to guide the model behaviour based on human-annotated gold standard solutions provided to the model, either prior to model training or through an active learning scheme. In Xin et al. (2020), a regression objective was introduced to align the model's representation of speaker identity with human-perceived speaker similarity, based on HitL feedback. Conventional speaker embeddings, such as *d*-vectors, don't always correspond well with perceptual judgements of speaker similarity (Saito et al., 2021). To address this, Xin et al. (2020) employed a perceptual similarity regression objective that regularises the latent speaker embedding space. This method was achieved by training a regressor on similarity scores provided by HitL participants. Their results indicate that the HitL-based regression objective improves both perceived naturalness and target speaker similarity. Saito et al. (2021) adopted a similar approach for modelling speakers, where they conditioned speaker identity generation on a *speaker similarity matrix*. This matrix was developed through an AL-based process, collecting speaker similarity judgements from HitL participants when required, as determined by the AL algorithm. The learned speaker representations correlate better with HitL-based speaker similarity ratings than conventional speaker identity representations.

8.2.1.2 Feature exploration

Many TTS models condition generation on high-dimensional representations to capture complex variation in speech. This includes, for example, the reference-based models (Part I), which jointly construct a latent space to sample from. HitL-centric *exploration* in latent representational spaces may yield perceptually-preferred representations (Harrison et al., 2020; van Rijn et al., 2021) or even representations previously unknown to the model (Udagawa et al., 2022). In principle, an exhaustive search through the latent space could be carried out by repeated sampling to identify a repre-

sentation that optimises a given perceptual objective. But latent representational spaces used in TTS are typically too large to support such a search. Methods like Markov Chain Monte Carlo with People (MCMCP) can be used instead, which progressively limit the exploration search space (Sanborn and Griffiths, 2007). In MCMCP, HitL participants are iteratively presented with two stimuli and asked to indicate preference regarding some perceptual objective. After each iteration, or *trial*, the two stimuli presented to the participants have been refined based on the preference indicated in the previous trial. One problem with this approach is that it is limited to binary responses; the HitL participants are only exposed to two stimuli at a time. This design greatly limits the amount of information that can be inferred from each HitL trial. Harrison et al. (2020) proposed Gibbs Sampling with People (GSP) instead. In GSP, HitL participants are presented with a control *slider*, allowing them to continuously adjust a single stimulus parameter at a time to maximise an evaluation criterion. This implementation allows more information to be conveyed per trial while limiting the stimulus space. Both MCMCP and GSP are general frameworks for HitL-centric exploration of feature spaces. Harrison et al. (2020) evaluated GSP for a diverse set of tasks, including one aimed at finding representations for emotive speech synthesis. They show that GSP converges faster than MCMCP, and yields perceptibly better results.

The stimulus space used for the emotive speech task in Harrison et al. (2020) was constrained to the manipulation of seven acoustic parameters. But GSP can also be employed for latent feature exploration, thus exploiting the expressive power of high-dimensional representations. But Harrison et al. (2020) only evaluated such latent GSP exploration for an image generation model. In a follow-up study (van Rijn et al., 2021), GSP was extended to explore speech representations modelled by a Global Style Token (GST) model (Wang et al., 2018). GST models, previously discussed in Section 3.3, encode corpus variation — unexplained by lexical content or speaker — into a small set of high-dimensional tokens. During inference, the influence of each token can be modulated through an attention weight parameter, thus dictating the token's relative contribution to the output. In van Rijn et al. (2021), HitL participants were tasked with discovering emotional representations, which were not necessarily represented in the training data. Here, the stimulus space was defined in terms of the token attention weights. HitL participants made iterative adjustments to the weights to maximise their individual perception of a particular emotion. In each trial, adjustments were limited to only one token attention weight, leaving all others fixed. Multiple cycles over all

attention weights were completed before terminating the process. Independent raters reliably identified the HitL-proposed emotional representations, and van Rijn et al. (2021) find that the representations generalise well across different target texts.

The chain of trials, performed in methods like GSP or MCMCP, can either be performed by a single HitL participant (*within-participant*) reflecting on the perception of just that particular participant, or by many different participants (*across-participant*). In van Rijn et al. (2021), each trial was completed by a different participant, reflecting on the shared perception of the group. They terminated the GSP process after cycling twenty times over each attention weight. While effective, this process is time-consuming, requiring up to 48 hours to generate forty samples (van Rijn et al., 2021). Presumably, across-participant methods yield more generalisable results, but alternative studies suggest that within-participant trials can still be effective for HitL applications in TTS.

Udagawa et al. (2022) use a process similar to GSP for speaker adaptation to an unseen speaker. They employed a HitL-centric linear search algorithm to find latent speaker representations resembling an unseen target speaker. Through this iterative process, they achieve results comparable to conventional speaker adaptation methods, without ever showing the TTS model a reference corresponding to the target speaker.

8.2.2 Acoustic feature control models

Models such as **FastSpeech 2** (Ren et al., 2020) are often described as *controllable* because they allow for changing the initially-predicted acoustic correlates of prosody. But many other types of TTS models can be said to be controllable. Reference-based control models can be controlled via the selection of a reference (Part I), prompt-based models can be controlled through a natural language description of the output (Part III) and a multi-speaker TTS model can be controlled through conditioning on a speaker label. But often, controllable TTS is used to refer to something more specific than that. So I use the term Acoustic Feature Control (AFC) model to refer to a model that:

1. separately predicts features that influence the speech generation process
2. can make predictions without explicit conditioning inputs
3. optionally allows for, through changes to the predicted features, generating different and diverse vocal renditions for the same input text.

AFC models could support model training via an IML process, in which a HitL-adjusted rendition is treated as a gold-standard training example. But this approach is rarely applied in practice. Here I focus on the design of AFC models in general. While all AFC models have the same core benefit, the ability to control aspects of the output, they may be designed differently to fit their specific use cases. Design differences are mainly reflected in (1) the underlying TTS architecture; (2) what *level* of control is employed; (3) which features can be controlled; and (4) how HitL participants interact with the system.

FastSpeech 2 (Ren et al., 2020) already provides rudimentary control over F_0, energy and duration. But, by default, predictions of these features are made on the frame level, which may be too high in resolution for conducive control of any kind. *FastPitch* (Łańcucki, 2021), a model similar to **FastSpeech 2**, predicts phone-level F_0 instead, which may offer a more intuitive control method. Alternative AFC models, which build on the **FastSpeech 2** architecture, have been proposed to abstract away from frame-level control. Cornille et al. (2022); Raitio et al. (2022a) suggested a hierarchical form of control, allowing for both phone- and utterance-level control, while Seshadri et al. (2022) proposed a word-level control scheme.

Cornille et al. (2022) proposed a *hybrid* approach to AFC, for control of expressive speech synthesis. Users first specified the overall speaking style by providing a reference utterance in the target style. The reference utterance conditions speech generation through the prediction of phone-level *prosody embeddings* which, among other information, captured prosodically salient features (duration, pitch, and loudness). A user could then make phone-level adjustments, via changes to the predicted prosody embeddings, to refine the prosodic rendition. As per my previous definition, this method could be considered a *hybrid* AFC method, since an initial reference is still required to control utterance-level prosodic variation. Raitio et al. (2022a) is similar to Cornille et al. (2022) but does away with any reference-conditioning. They predict both utterance-level and phone-level features, but instead of latent representations, control is achieved through changes to interpretable features. The model first predicts global F_0, duration and energy features based on the target text alone, which the user can then modulate. Both approaches presented by Cornille et al. (2022) and Raitio et al. (2022a) are *hierarchical*: the chosen utterance-level values influence the phone-level feature values, which the user can further refine. In Raitio et al. (2022a), each of these features can be controlled independently, while in Cornille et al. (2022) there

is no such separation of features. Seshadri et al. (2022) takes the middle ground by predicting both latent and disentangled features to support the control of word-level emphasis. First, latent word-level emphasis features are predicted from the target text. Then, these latent representations influence the prediction of phone-level F_0, energy and duration. Emphasis control is achieved by changing individual word-level emphasis representations initially predicted by the model.

Models that otherwise provide no control of prosody, such as **Tacotron** (Wang et al., 2017), can also be extended into AFC models (Shechtman et al., 2021; Mohan et al., 2021; Raitio et al., 2020). Similar to Seshadri et al. (2022), Shechtman et al. (2021) proposed a word-level scheme for control of lexical focus. Mohan et al. (2021) aimed to provide a general control method that is interpretable, disentangled, and temporally fine-grained. They modified a **Tacotron** model with phone-level control of F_0, energy and duration to enable targeted modifications to separate acoustic features. However, they suggested that a higher level of abstraction may be more suitable for HitL applications. Raitio et al. (2020) aims to provide controls for generating speech according to different speaking styles. During training, the model learns to predict utterance-level F_0 mean and range, energy, duration, and spectral tilt from the target utterance acoustics. They use teacher forcing, so during inference, these can be predicted from the target text alone. Control is achieved through biasing these five predicted features.

Although these AFC models differ in several ways, they all follow the same principle: enabling a user to specify a vocal rendition for the given target text. Typical TTS models only predict a single rendition for a given target text. But HitL-suggested changes to acoustic variance allow for varied renditions (Mohan et al., 2021), diverse expressions and speaking styles (Raitio et al., 2022a) which can improve the perceived naturalness (Shechtman et al., 2021; Mohan et al., 2021; Raitio et al., 2022a) or appropriateness (Seshadri et al., 2022) of the generated speech.

Chapter 9

Controllable speaking styles using a large language model

Text-To-Speech (TTS) models such as **FastSpeech 2** (Ren et al., 2020) and FastPitch (Łańcucki, 2021) explicitly predict acoustic correlates of prosody. These models offer interpretable prosody modification: during inference, the predicted values of these acoustic correlates could, in principle, be tuned by a human expert to fulfil a particular prosodic requirement.

The current chapter presents *Controllable Speaking Styles Using a Large Language Model* (Sigurgeirsson and King, 2024), which proposed an acoustic feature control scheme for a **FastSpeech 2** (Ren et al., 2020) model. Many of the Acoustic Feature Control (AFC) models listed in Section 8.2 extend the **FastSpeech 2** architecture to achieve a particular form of control. A primary objective in the current study was to evaluate whether **FastSpeech 2** supports prosody control without any such extensions. Typically, controlling a TTS model in this way would require a Human-in-the-Loop (HitL) participant. However, such HitL-based approaches are time-consuming, as evidenced in Section 8.2.1.2. This study proposed instead to employ a Large Language Model (LLM) to take the place of the participant.

9.1 Research objective

LLMs have shown excellent performance in various language-related tasks. Given only a natural language query text (the *prompt*), such models can be used to solve specific, context-dependent tasks. As discussed later in Part III, many recent TTS models

aim to provide such prompt-based control for guiding speech generation. Such methods can predict varied and plausible renditions for the same input text by conditioning on different prompts. The aspects of speech that can be controlled vary between models, but an example description prompt could be: *"The man speaks slowly, he sounds energetic and his voice is high-pitched"*. These models do, therefore, offer very flexible and user-friendly means to control. However, existing methods require a vast, prompt-labelled speech corpus to train a prompt-conditioned encoder.

In contrast, I instead employed a frozen LLM to directly suggest prosodic modifications for the TTS model, using contextual information provided in the prompt. Therefore, the proposed method does not require training of a prompt encoder or a style-labelled corpus. Given the prompt, the LLM suggests which acoustic features to change and how to change them. These changes are then applied to produce a new, prosodic rendition. *InstructGPT* — an LLM that has already been fine-tuned to follow natural language instructions (Ouyang et al., 2022) — was used in the current work. The prompt can be designed for a multitude of tasks, as the flexibility of the design allows for arbitrarily changing the objective that the LLM is tasked with.

The underlying TTS model used in the current study is based on the **FastSpeech 2** (Ren et al., 2020) architecture. The LLM is prompted to suggest modifications to the acoustic features predicted by the model, based on the target text and optional contextual clues. To the best of my knowledge, the proposed method was novel and the first to prompt an LLM *using only natural language* to solve a specific task in TTS. To investigate the feasibility of this approach, I asked:

> **RQ 9-1:** *Can a large language model be instructed to modify acoustic features to improve expressive speech synthesis?*

Three different tasks were evaluated in the current study:

1. **Neutral discourse**: where the LLM is not provided with any contextual clues that may inform it of an appropriate prosodic rendition,
2. **Speaking style generation**: where the LLM is prompted with a target speaking style described using natural language,
3. **Expressive conversational speech**: where the LLM predicts the appropriate prosodic rendition based on the previous line in an expressive dialogue.

A considerable contribution in the current study was the more interpretable approach to steering TTS models through natural language prompting: the suggestions made by the LLM can be explained in terms of known acoustic features. But beyond this contribution, I was interested in evaluating whether models like **FastSpeech 2** even allow for such direct model interaction in the first place:

> **RQ 9-2:** *Can the acoustic features, predicted by models like **FastSpeech 2**, be effectively modified to perceptually improve expressive speech generated by such models?*

Two different perceptual qualities were investigated: perceived naturalness and perceived *appropriateness*. The proposed method employed a control scheme where the predicted values of F_0, energy and duration can be changed to more suitable ones, given one of the three task contexts previously described. Adequate adjustments to the predicted acoustic features could allow for improving the synthesised prosodic rendition, therefore improving the perceived overall naturalness:

> **H9-1a:** *Speech generated by the proposed model is perceived as more natural than speech generated using originally predicted values in the **neutral discourse** task*

In addition to a baseline model, which employed unadjusted acoustic features, I compared the proposed method to two reference-based control models: a Prosody Transfer (PT) model and a Global Style Token (GST) model. These two alternative models had direct access to the target output: the reference used for conditioning the model was a ground truth rendition of the target text. As the reference contains a ground truth prosodic rendition for the target text, these two *oracle* models were expected to generate more natural speech than the baseline model. But, I believed that the proposed control method could yield alternative, yet plausible, expressive prosodic renditions:

> **H9-1b:** *Speech generated by the proposed model is as natural as speech generated by the reference-based models in the **neutral discourse** task*

The proposed method was additionally evaluated in two expressive speech tasks, **speaking style generation** and **expressive conversational speech**. A baseline **FastSpeech 2** model learns a single acoustic mapping for a given input text. So, without additional

conditioning inputs, a model trained on inexpressive data naturally learns to map text to inexpressive speech. Such a model may therefore be perceived as inappropriate when expressive speech is anticipated. I hypothesised that the proposed control scheme allows for changing an inexpressive rendition into an expressive one:

> **H9-1c:** *Speech generated by the proposed model is perceived as more appropriate than speech generated by* ***FastSpeech 2*** *for expressive TTS.*

Listeners might be biased in favour of *any* acoustic variance when listening to utterances whose propositional content indicates an expressive speaking style. But I hypothesised that the predicted acoustic features (F_0, energy, and duration) have to be modified in a way that makes sense for the target text and other contextual information provided, to result in improved appropriateness. Therefore:

> **H9-1d:** *Speech generated by the proposed model is perceived as more appropriate than speech corresponding to pseudo-random adjustments to acoustic features*

9.2 Model architecture

In this work, a slightly modified **FastSpeech 2** (Ren et al., 2020) was employed as the **baseline** TTS architecture. **FastSpeech 2** has been described in Section 2.3.2.3, but briefly, **FastSpeech 2** comprises a phoneme encoder, a variance adaptor, and a mel-spectrogram decoder. The encoder and decoder each comprise four feed-forward transformer blocks (Vaswani et al., 2017). In **FastSpeech 2**, the variance adaptor predicts F_0 and energy (after predicting each phone's duration). Energy prediction is dependent on the predicted value of F_0. I wished to modify both, choosing to do this per word, whereas **FastSpeech 2** predicts per frame. Therefore, the **FastSpeech 2** variance adaptor and duration model were replaced with the *low-level prosody predictor* and Gaussian upsampling module from Zaïdi et al. (2022).

The low-level prosody predictor produces phone durations and per-phone speaker-normalised log-F_0 and log-energy. By its design, the baseline model architecture predicts these features jointly, which can then be modified separately. The Gaussian upsampling module predicts a duration distribution for each phone based on the encoder output and the previously predicted log-F_0, log-energy and duration. The module then samples a phone duration from each of those distributions. Per-phone log-F_0 and log-energy predictions are projected and summed with the encoder output, then upsampled

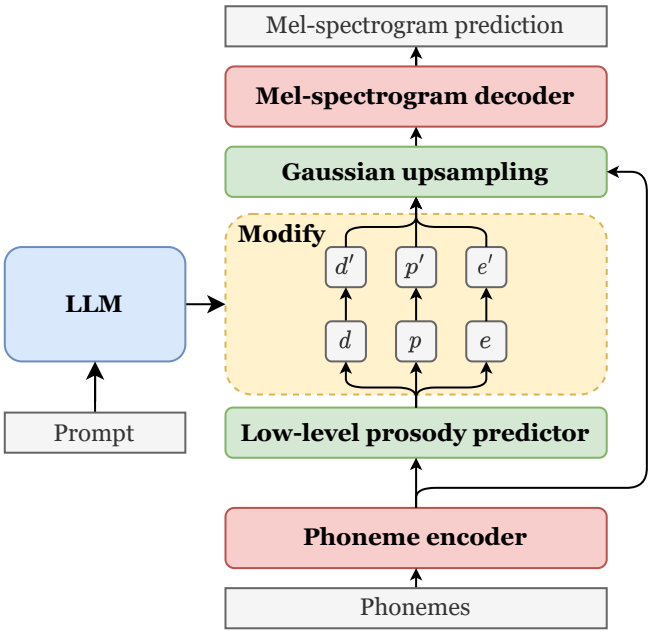


Figure 9.1: During inference, the low-level prosody predictor predicts per-phone duration, log-F_0, and log-energy sequences (**d**, **p**, and **e**, respectively). Those initial sequences are then modified (into $\mathbf{d}'$, $\mathbf{p}'$, and $\mathbf{e}'$) based on the suggestion made by the LLM, before continuing the forward pass.

using these phone durations. The decoder then predicts mel-spectrogram frames from this upsampled latent representation. During training, the model uses ground-truth values for duration, log-F_0, and log-energy. The proposed method utilises this baseline architecture and the same model weights, but makes inference-time modifications to the predicted phone-level acoustic features, as illustrated in Figure 9.1.

9.3 Modification method

Given the sequences of per-phone durations (**d**), log-F_0 (**p**), and log-energy (**e**), predicted by the low-level prosody predictor[1], for an input sequence of length T:

$$\mathbf{d}=d_1,\ldots,d_T, \quad \mathbf{p}=p_1\ldots p_T, \quad \mathbf{e}=e_1\ldots e_T$$

predicted by the TTS model, the aim is to find more appropriate values

$$\mathbf{d}'=d'_1,\ldots,d'_T, \quad \mathbf{p}'=p'_1\ldots p'_T, \quad \mathbf{e}'=e'_1\ldots e'_T$$

[1]Hence all value sequences are of length T as upsampling of log-energy and log-F_0 is performed after they are modified.

Hypothetically, every phone-level feature could be precisely adjusted towards any particular perceptual objective. But, importantly, the task has to be simple enough for an LLM to be able to solve it. So, even if sophisticated, high-precision control schemes would be possible in theory, I suggested a simpler 2-level modification procedure instead. The procedure allows for controlling both utterance ('global') and word ('local') prosodic effects. Global modifications may convey, for example, some aspects of emotion, attitude or speaking style. Local modifications may, for example, be used for lexical stress. The method does not directly model any particular speaking styles, nor does it require data annotated with them.

Global modifications are applied to all phones in the utterance using three coefficients. Durations and log-energies are scaled by G_d and G_e, respectively. Global log-F_0 scaling resulted in artefacts, so instead log-F_0 values are shifted by G_p. G_d and G_e are limited to $[0.5, 2]$ while G_p is limited such that p'_i remains within the natural F_0 range of the target speaker, determined based on corpus statistics. For an input text consisting of W words, local (word) modifications are realised using three coefficient sequences

$$\delta_1, \ldots, \delta_W, \quad \pi_1, \ldots, \pi_W \quad \text{and} \quad \varepsilon_1, \ldots, \varepsilon_W$$

The values resulting from both global and local modifications for phone i appearing in word j are:

$$d'_i = d_i \cdot G_d \cdot \delta_j, \qquad G_d \in [0.5, 2], \quad \delta_j \in [1.0, 2.0] \tag{9.1}$$

$$e'_i = e_i \cdot G_e \cdot \varepsilon_j, \qquad G_e \in [0.5, 2], \quad \varepsilon_j \in [1.0, 2.0] \tag{9.2}$$

$$p'_i = p_i + G_p + \pi_j, \qquad p_i + G_p + \pi_j \in [p_{\min}, p_{\max}] \tag{9.3}$$

where $p_{\min}$ and $p_{\max}$ are the minimum and maximum allowed changes in F_0 determined by the corpus statistics for that phone. The low-level prosody predictor predicts log-scale and speaker-normalised energy and F_0. So, before applying the modifications in Equations 9.2 and 9.3, they are first converted to linear scale and de-normalised. After applying the modifications, they are re-normalised and converted back to a log scale before being passed to the Gaussian upsampling module. No changes are made to F_0 or energy of unvoiced phones, nor are pauses modified (changing pause durations resulted in considerable artefacts, possibly caused by inaccurate alignments).

9.4 Prompt construction

I used InstructGPT (Ouyang et al., 2022) via the OpenAI API[2]. InstructGPT is a fine-tuned version of GPT-3 (Brown et al., 2020), a 175 billion parameter autoregressive language model trained on over 400 billion tokens. InstructGPT has been fine-tuned using a mixture of supervised training and reinforcement learning to follow instructions supplied to it through a natural language prompt (Ouyang et al., 2022). In the current work, the goal of the LLM is to generate appropriate values for the global and local coefficients given the target text. For this task, the prompt consists minimally of a description of the task and the target text. However, using minimal prompts resulted in inconsistent and nonsensical responses. Here, I explain which text-based instructions were included in the prompt to get more consistent predictions from the model. Snippets from the prompt are included throughout this section, and the full prompt is provided in Appendix E.

Adhering to the prompt

Much effort was required to get InstructGPT to give a *sensical* solution to the proposed task. It would frequently hallucinate new words to modify, solve a completely different task, confuse parameters and their ranges, or fail to provide a comprehensive answer at all. Likely, newer LLMs do not need as much guidance to produce a sensical response.

The ranges from which global and local coefficients are drawn, shown in Equations 9.1-9.3, are not the same. The LLM struggled with consistently predicting values within an appropriate range for each coefficient. Because of this, the LLM was instructed to predict global values in the range $[-5,5]$ and local values in $[0,5]$. These values were then linearly mapped to the appropriate range for each coefficient. The instruction prompt explained what each coefficient controls and what a negative or positive value represents.

Prompt Snippet - Task Explanation

```
...So for each of those attributes, tell me how much to change
them: (0: the standard value, -5: the minimum value, 5: the
maximum value). A positive value for duration means a slower
speaking rate, a negative value means a faster speaking rate...
```

[2]https://openai.com

A mixture of few-shot prompting (Min et al., 2022) — where the instructional prompt includes gold-standard examples of how to solve the task for particular inputs — and chain-of-thought (Wei et al., 2022) prompting — where the prompt dictates the LLM to motivate intermediate outputs which affect downstream predictions — was also helpful in this task. Human-generated solutions were included in the prompt (few-shot), where each intermediate step was reasoned (chain-of-thought). Ten such example solutions were included in the instruction prompt.

Prompt Snippet - A Few Shot Example

The instruction prompt includes 10 gold-standard few-shot examples, like the one shown below:

Target speaking style: *Inspirational speech*
Target text: *'You can do it!'*
Here is how I would change the prosody:

Word	Prominence
You	1
can	3
do	0
it	1

Pitch	Energy	Duration
2	3	-2

Explanation: *We shorten the overall duration since the speaker is probably excited, given the target speaking style. We further raise the pitch and energy for a more excited-sounding voice, which is fitting for this target text. We add a low-level prominence to the words 'You' and 'it'. The word 'can' is the most important word in the sentence since the speaker is likely convincing the listener that they can do something.*

An additional set of rules that the LLM was instructed to follow was also included. For example, the model is told to predict the parameters independently of the target voice. Without such a specification, the LLM would frequently base predictions on a hypothetical speaker identity never mentioned in the instruction prompt. The LLM would often skip words from the target text in its prediction, and often make up new ones instead. But, giving the LLM strict formatting instructions about how to produce a solution mitigated this issue. Together, the task description, the few-shot examples, the rules, and the formatting instructions form the prompt that is supplied to the LLM *only in the form of natural language*.

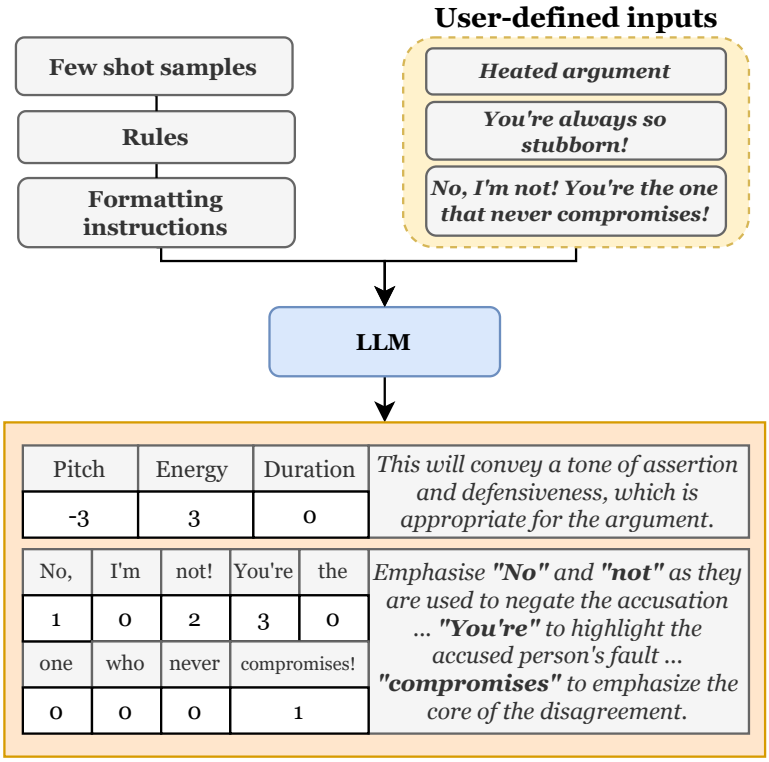


Figure 9.2: The LLM suggests acoustic modifications, given the target text and, optionally, a speaking style or previous dialogue context. The rules, supplied to the LLM in the prompt, ask the LLM to produce reasoning for the modifications.

Prompt Snippet - Formatting Instructions

```
...Report the sentence-level change of pitch, energy and duration
in a separate table.  Pitch, energy and duration should be
columns in the table.  Pitch first, then energy and finally the
duration.

Report the prominence level of each word in a separate table.
Remove any punctuation, as we do not have to predict prominence
for those symbols.

Here are templates for the two tables:  ...
```

The complete procedure is illustrated in Figure 9.2. The flexibility of the method means that the LLM can be prompted to make predictions based just on text or with additional contextual information such as speaking style or dialogue. The instruction prompts used in the experiments included clear instructions on which task the LLM should solve.

9.5 Experimental setup

I compared the **proposed** method to several other models as described in Sections 9.5.3-9.5.4. But, the slightly modified **FastSpeech 2** model, as described in Section 9.2, serves as the **baseline** comparison model for all three tasks.

9.5.1 The tasks

The first task is **neutral discourse**. Here, the LLM receives no contextual information, beyond the target text, and is prompted to only make local modifications to appropriately emphasise words in the text. Such prompting does not yield much more expressive speech than what is reflected in the training corpus. But it was hypothesised that the LLM would predict sensible emphasis variation outside the range of a typical **FastSpeech 2** model, thus leading to improved perceived naturalness.

In the second task, **speaking style generation**, the model is prompted with a description of a target speaking style. Here, the model predicts both local and global adjustments. A small text corpus was created to evaluate the model's performance in this task. First, a number of *speaking styles* were selected and, for each one, several target texts that were deemed appropriate for that style were assigned. The flexibility of this approach means that we can define highly specific speaking styles for evaluation. The list included styles such as *"frightened"*, *"in a hurry"* and *"speaking to a child"*. They were chosen solely on the basis that they likely correspond with different settings of the modification coefficients.

Lastly, in the **expressive conversational speech** task, the model is prompted to predict both local and global adjustments for an expressive dialogue. The model is told the previous line in the dialogue and the target text, *not* a target style. First, a list of *hidden* speaking styles, which were not provided in the instruction prompt, was created. Then, several two-line dialogues were written, where the target text would naturally fit the speaking style.

The speaking styles employed for this task are later shown in Figure 9.5 (p. 123). These include *"Apologetic"*, *"Friendly"* and *"Heated argument"* for example. An actual prediction for *"Heated argument"* was previously shown in Figure 9.2 (p. 115) as well as the example dialogue used for that sample.

An Example of Expressive Dialogue

For each *hidden* speaking style label, I created 10 evaluation dialogues. This is an example for the *sarcastic* label:

Hidden style: *The speaker is sarcastically responding to someone*

Previous line: *I think we should take a break and come back to this later.*

Target text: *Oh sure, because procrastination is always the best strategy.*

9.5.2 Model training

The proposed model and all other evaluated models were trained on LJ Speech (Ito and Johnson, 2017). The LJ Speech corpus is single-speaker and not designed specifically for expressive TTS training. Utterances shorter than 1.5 s were removed, leaving approximately 11,000 utterances for training. 80 bin mel-spectrograms are extracted from 22 050 Hz waveforms. Phone alignments were found using the Montreal Forced Aligner (McAuliffe et al., 2017). F_0 is estimated using REAPER[3] and the l^2-norm of spectrogram frames was extracted for energy.

Each model was trained for 24 hours on 8 NVIDIA A100-SXM-80GB GPUs with a batch size of 192. A pre-trained HiFi-GAN vocoder Kong et al. (2020) was fine-tuned on mel-spectrograms generated by each model trained, thus creating a matched vocoder for each model.

9.5.3 Evaluation of naturalness

The perceived naturalness of the **proposed** method in the **neutral discourse** task was measured to test **H9-1a**. The **proposed** method was expected to be perceived as more natural than the **baseline** model, which employs unadjusted acoustic features (**H9-1a**). The **proposed** method was also compared to the two reference-based **oracles**. **Daft-Exprt** (Zaïdi et al., 2022) was employed as a representative PT model since its architecture is derived from **FastSpeech 2**. A GST model was also evaluated, in which the style token layer from Wang et al. (2018) is added to the reference encoder in **Daft-Exprt**. I performed inference with the style-token model using reference conditioning as described in Wang et al. (2018). The models were denoted as $\textbf{Oracle}_{\textbf{PT}}$ and $\textbf{Oracle}_{\textbf{GST}}$, since they have full access to the ground truth target mel-spectrogram. It

[3] https://github.com/google/REAPER

was previously predicted that the **proposed** method would be perceived as comparably natural to these two oracles (**H9-1b**).

Perceived naturalness was evaluated with a standard Mean Opinion Score (MOS) design. Native English-speaking listeners were recruited via Prolific [4], and eight different listeners evaluated each sample. 30 unseen utterances from LJ Speech were evaluated, resulting in $30 \times 8 = 240$ ratings per model. These results are compared against the perceived naturalness of the 30 **ground truth** utterances — used to create the model samples — to account for any model-based degradation in quality.

9.5.4 Evaluation of appropriateness

Hypotheses **H9-1c** and **H9-1d** are evaluated in a listening test involving the two expressive tasks, **speaking style generation** and **expressive conversational speech**. Here, the **proposed** method was compared to the **baseline** and **shuffle**. For **speaking style generation**, seven different speaking styles were created, and 10 target texts were assigned to each one. Perceived appropriateness was evaluated with a preference A/B/C design, yielding $7\times10\times8$=560 sets of stimuli. Participants were shown the target text and speaking style when performing this evaluation. Participants were instructed to judge appropriateness based on how well they thought the vocal rendition fit the target text. A similar strategy was employed for the **expressive conversational speech** task. Six *hidden* speaking styles were created, each comprising 10 two-line dialogues corresponding to the style, resulting in $6\times10\times8$=480 sets of stimuli in total. Participants first read the previous dialogue line before choosing the most appropriate rendition, out of **proposed**, **baseline**, and **shuffle**, using the same A/B/C preference design.

9.6 Results

9.6.1 The neutral discourse task (H9-1a and H9-1b)

Naturalness results are reported in Table 9.1, alongside judgements for **ground truth** utterances. A Linear Mixed-Effects (LME) model was used to evaluate differences in mean MOS ratings between the four evaluated models and **ground-truth** utterances, which are also denoted as `systems` in the following analysis. The LME model included `system` as a fixed effect and random intercepts for `rater` and `utteranceID`

[4] https://www.prolific.co

to account for repeated measures. **Ground truth** was used as the reference condition `system`. Repeated Holm-adjusted (Holm, 1979) t-tests were employed to perform pairwise comparisons between `systems`.

Table 9.1: MOS results for the **neutral discourse** task. 95% confidence intervals estimated using bootstrapping (10,000 resamples).

Model	Naturalness MOS
Baseline	3.06 ± 0.17
Proposed	3.06 ± 0.16
$\text{Oracle}_{\text{GST}}$	3.20 ± 0.16
$\text{Oracle}_{\text{PT}}$	3.47 ± 0.16
Ground truth	4.14 ± 0.13

All four TTS `systems` were associated with a mean decrease in MOS ratings compared to **ground truth**. All these differences were significant ($p \leq .001$, $\alpha = .05$), suggesting an overall modelling degradation in perceived naturalness.

No statistical difference between **baseline** and **proposed** was revealed in their pairwise comparison ($T = 0.44$, $p = .66$, $p_{\text{adj.}} = .66$), indicating that the **proposed** condition did not have the predicted improving effect in terms of naturalness:

> **H9-1a**: *Speech generated by the proposed model is perceived as more natural than speech generated using originally predicted values in the **neutral discourse** task* **Reject ✗**

There was no significant difference in naturalness ratings between **proposed** and $\textbf{oracle}_{\textbf{GST}}$ ($T = 1.33$, $p = .19$, $p\text{adj.} = .48$). However, $\textbf{oracle}_{\textbf{PT}}$ received significantly higher ratings than **proposed**, indicating a minor but statistically significant difference ($T = 3.07$, $p = .005$, $p_{\text{adj.}} = .02$). Based on these mixed results:

> **H9-1b**: *Speech generated by the proposed model is as natural as speech generated by the reference-based models in the **neutral discourse** task* **Reject ✗**

This task represented the optimal case for the two **oracles** since an appropriate reference could be used for conditioning; namely, same-text ground-truth utterances spoken by the training voice. In this condition, $\textbf{oracle}_{\textbf{PT}}$ was rated better than all other `systems`, including the **proposed** `system`. I take the results presented in Table 9.1 as an indication that, under this optimal condition, guiding prosody with a reference utterance is indeed better than the method proposed. However, synthesising particu-

lar speaking styles or contextually-appropriate prosody using reference-based models requires finding an appropriate reference for conditioning. Finding such a reference is a non-trivial task when the speech corpus is not style-labelled. Furthermore, most speech corpora are limited in the types of speaking styles that the voice performs. These shortcomings eliminate the **oracles** from the two other tasks covered in Sections 9.6.2-9.6.3; underlining the limitations of reference-based models.

9.6.1.1 Relations to variation in fundamental frequency

F_0 variation in `systems` was further studied to reflect on the naturalness results. To limit assumptions about the data, a Kruskal-Wallis H-test (Kruskal and Wallis, 1952) was conducted to test for a significant effect of `system` on F_0 variation. Given a significant effect, Mann-Whitney U-tests (Mann and Whitney, 1947) were then conducted for pairwise comparisons between `systems`.

There was a significant effect of `system` on F_0 variation ($\chi^2(4) = 45.89$, $p < .001$). The two **oracle** `systems` exhibit F_0 variation that was comparable to **ground truth** utterances ($U \geq 286.0$, $p \geq 0.09$). However, both the **baseline** `system` and the **proposed** `system` were significantly less varied than **ground truth** utterances ($U \geq 413.0$, $p \leq 1.4e-06$). The **proposed** model was the only model statistically comparable to the **baseline** ($U = 190.0$, $p = .45$), suggesting that the F_0 variance present in the **oracle** models — derived from the **ground truth** references — had an improving effect that the **proposed** model did not replicate.

9.6.2 Speaking style generation (H9-1c and H9-1d)

Results from the **speaking style generation** task are broken down in Figure 9.4 and summarised in Table 9.2. A hierarchical Bayesian mixed effects model estimated `system` (**proposed**, **baseline**, and **random**) appropriateness preference probabilities. The model included random intercepts to account for repeated measures and `rater`-level variation. Pairwise comparisons were made based on credible difference by analysing Highest Density Intervals (HDIs) of posterior preference distributions. **Random** was used as the reference `system`.

The analysis of posterior intercepts revealed that both **baseline** and **proposed** were preferred at a credibly higher rate than **random**. This difference is more apparent for **proposed** (95% HDI: $[0.823, 1.277]$) than **baseline** (HDI: $[0.086, 0.6]$). However,

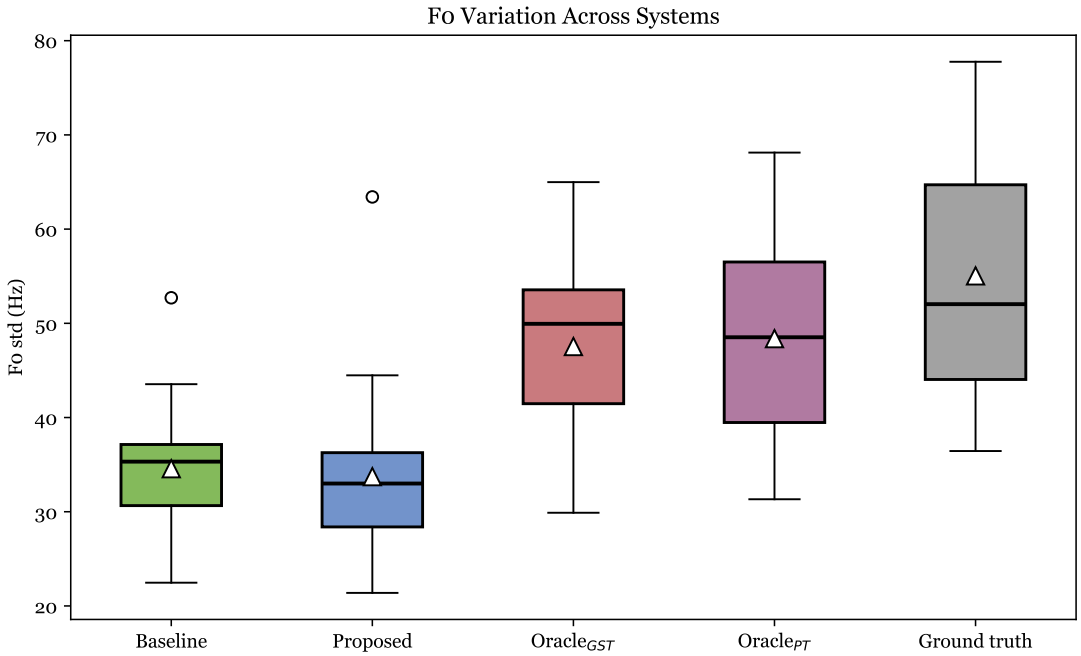


Figure 9.3: Distribution of F_0 variation (standard deviation in Hz) across utterances for each system in the **neutral discourse** task. Each box represents the interquartile range, with the median (horizontal line), mean (triangle), and statistical outliers (circle).

as Figure 9.4 illustrates, this did not hold for all speaking styles used in the survey. In the majority of cases, the proposed `system` was rated as most appropriate and was preferred at a >70% higher overall rate than **baseline**; a credible difference (95% HDI: [0.516, 0.945]). Raters did not indicate a high preference for **random**, suggesting that random variation alone was insufficient to bias preference rates.

Table 9.2: A/B/C appropriateness preference results for both the target style and dialogue tasks, and overall results across both tasks.

Task	Proposed	Baseline	Random
Speaking style generation	51.4%	30.9%	17.7%
Expressive conversational speech	48.4%	31.0%	20.6%
Overall	49.9%	31.0%	19.1%

9.6.3 Expressive conversational speech (H9-1c and H9-1d)

Results for the **expressive conversational speech** task are summarised in Table 9.2 and broken down by hidden speaking style in Figure 9.5. Another Bayesian mixed-effect model was employed to evaluate credible differences between system preference rates. Again, `rater` and `utteranceID` are included as random effects.

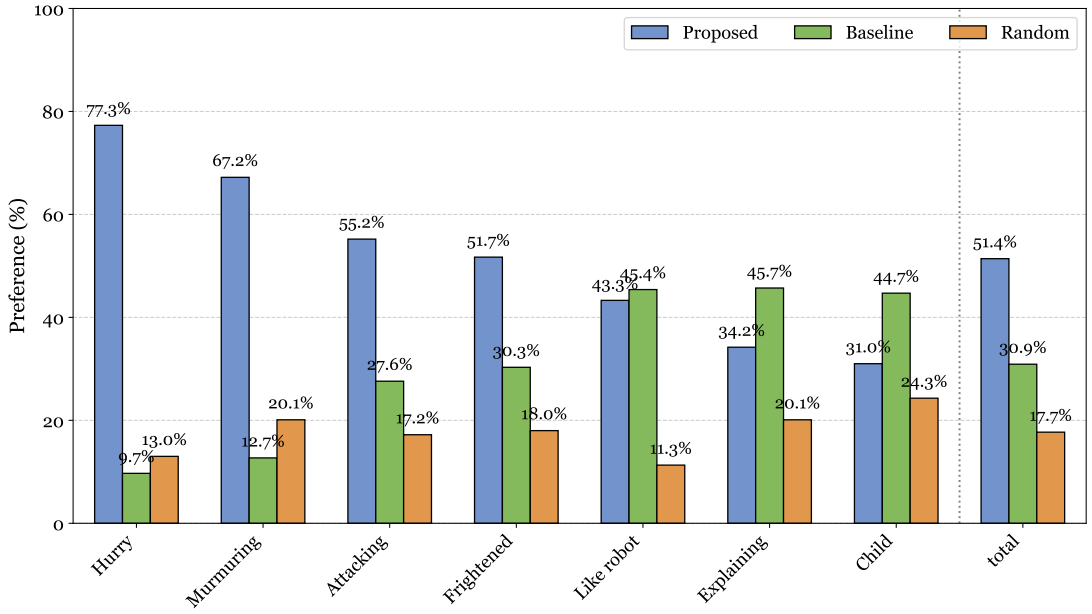


Figure 9.4: Preference results for the **speaking style generation** task

Based on the analysis of the posterior preference distributions, both **baseline** (95% HDI: $[0.133, 0.647]$) and **proposed** (95% HDI: $[0.619, 1.084]$) are preferred over **random** at a credible rate. A follow-up pairwise comparison showed that the mean preference for **proposed** was credibly higher than for **proposed** (95% HDI: $[0.217, 0.647]$).

Results in Figure 9.5 show that the proposed method is better suited for more expressive styles, such as *heated argument* or *excited*, but less so for more neutral styles such as *formal discussion*. And this was particularly clear for the *sarcastic remark* hidden speaking style. Overall, however, participants indicated a clear overall preference for the proposed method when compared to either the **baseline** or **random**. This outcome matches the previously reported results for the **speaking style generation** task. Based on the combined results, both hypotheses were accepted:

> **H9-1c**: *Speech generated by the proposed model is perceived as more appropriate than speech generated by **FastSpeech 2** for expressive TTS.* **Accept** ✔
>
> **H9-1d**: *Speech generated by the proposed model is perceived as more appropriate than speech corresponding to pseudo-random adjustments to acoustic features* **Accept** ✔

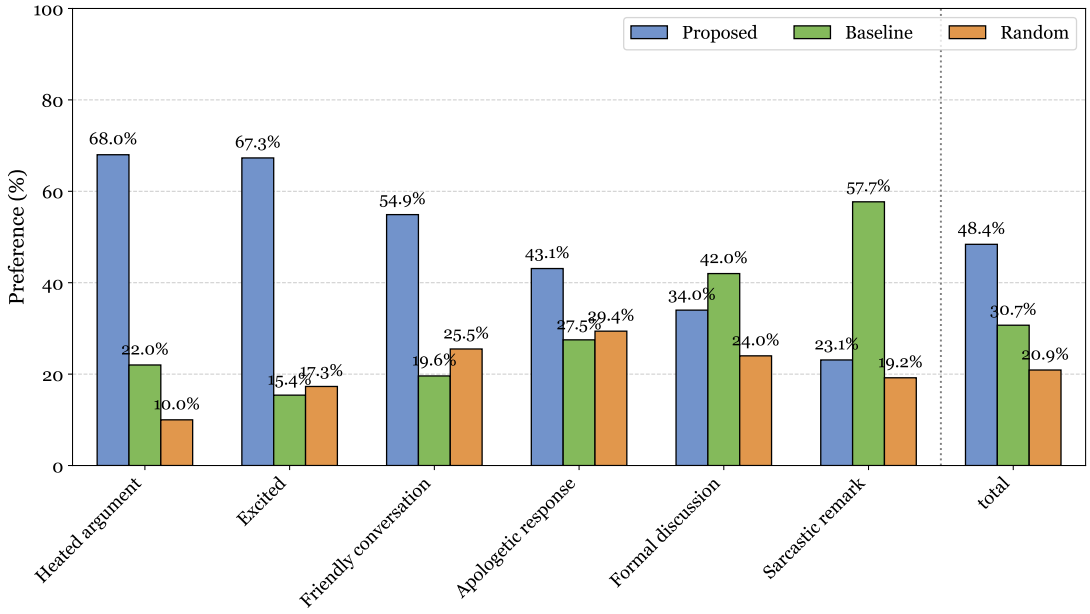


Figure 9.5: Preference results for the dialogue task.

The Sarcastic LJ

The results for the baseline model reflect how people perceive an unaltered LJ. And like the proposed model, the salience of this perception varies across categories. The results indicate that LJ is particularly appropriate when the prepositional content calls for a sarcastic tone (see Figure 9.5).

9.6.3.1 Relations to variation in fundamental frequency

Correspondence between preference rates and F_0 variation across `systems` in the *expressive conversational speech* task was additionally analysed. F_0 variation differs significantly between `systems`, based on a Kruskal-Wallis H-test ($\chi^2(2) = 10.60$, $p = .004$). The **proposed** method was marginally more varied in F_0 than the **baseline** model according to a follow-up pairwise test ($U = 2349.0$, $p = .002$). Increased variation in F_0 may have improved the appropriateness of the **proposed** model where expressive speech was anticipated. The same analysis showed that the F_0 variation of the **random** `system` did not reliably differ from that of the **proposed** model ($U = 2197.0$, $p = .023$). This result indicated that *any* F_0 variation alone is not sufficient to enhance perceived appropriateness; this also depends on the manner in which it varies.

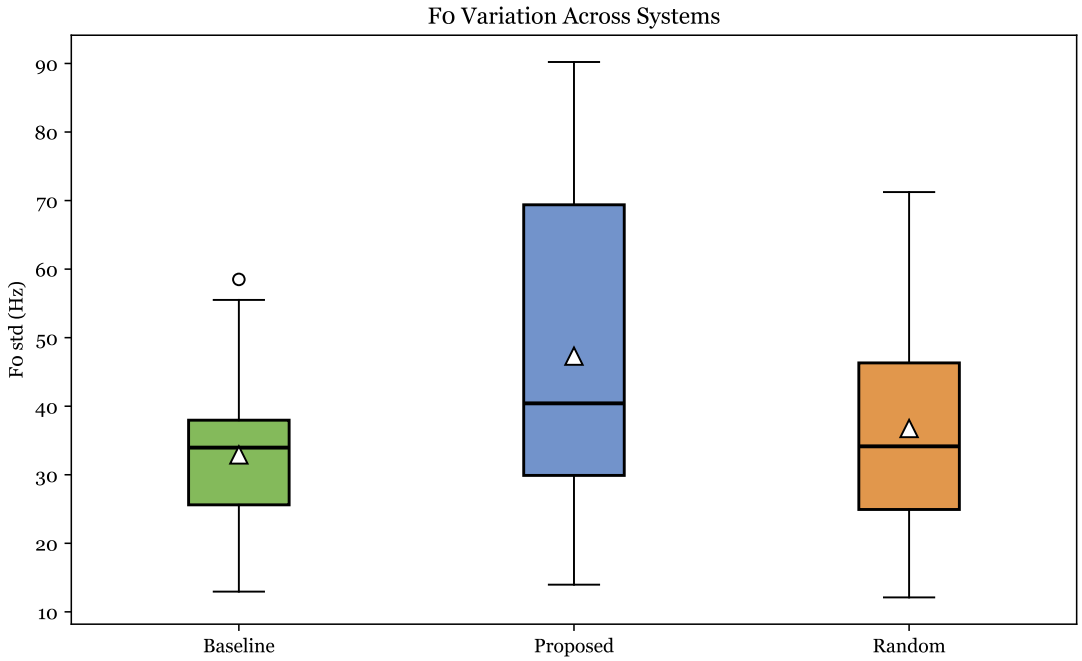


Figure 9.6: Distribution of F_0 variation (standard deviation in Hz) across utterances for the **expressive conversational speech** task.

9.7 Conclusion

Previously, I asked: **RQ 9-2:** ***Can the acoustic features, predicted by models like FastSpeech 2, be effectively modified to perceptually improve expressive speech generated by such models?***. The results demonstrate that the **proposed** method can adjust the vocal rendition to improve listeners' perception of novel speaking styles (i.e., not represented in the training data). But, compared to a **baseline** model employing unadjusted acoustic features, the method did not improve perceived naturalness. As discussed in Section 9.3, the proposed control scheme had to be simple enough for the LLM to be able to solve the control task. The task simplification may have inhibited the **proposed** model in making appropriate adjustments for the **neutral discourse** task, where more nuanced adjustments may be required. Yet, the **proposed** model was perceived as more appropriate than the **baseline** for expressive speech synthesis. Across the two expressive speech tasks, the **proposed** model was rated most appropriate in 50% of cases vs. 31% for a **baseline** model.

The results also illustrate the impact of training data selection on the performance of the **proposed** method. The training corpus used in this study, LJ Speech, lacks expressive speech. So, naturally, it is expected that the **baseline** model — which predicts acoustic variation based solely on the LJ Speech corpus statistics — would perform

better for neutral, narrative purposes than for expressive TTS. This is reflected in the results where participants rated the **proposed** model as less appropriate, compared to the **baseline** model, for less expressive speaking styles (such as *"formal discussion"* and *"explaining"*). In contrast, the **proposed** method was favoured in contexts where expressive speech is expected. These results suggest that, despite the control scheme's simplicity and the limitations of the training data, the **proposed** method can still enhance perceived appropriateness in expressive scenarios.

I also asked: **RQ 9-1:** ***Can a large language model be instructed to modify acoustic features to improve expressive speech synthesis?***. The results indicate that InstructGPT can be prompted to control TTS models for expressive applications. Results in Section 9.6.3 demonstrate that the LLM made appropriate adjustments to predicted acoustic features. Meanwhile, a model that employed randomly-varied acoustic features did not improve perceived appropriateness. In the current work, the LLM stands in for a HitL participant to steer the model. While the control scheme demonstrated potential for steering a TTS model, it does not provide sufficient insights into how *usable* the proposed control method actually is. This question is tackled in the study presented in the next chapter.

Chapter 10

Prosody transfer with a human-in-the-loop

In *A Human-in-the-Loop Approach to Improving Cross-Text Prosody Transfer* (Maurya and Sigurgeirsson, 2024), we investigated whether Human-in-the-Loop (HitL) feedback can be used to improve the prosody representations modelled by a typical Prosody Transfer (PT) model. To support this investigation, we adopted the Acoustic Feature Control (AFC) scheme proposed in Sigurgeirsson and King (2024) and presented in Chapter 9. Our investigation also examined the usability of this approach. We analysed the effort required to complete the task and the self-reported success rate of HitL participants. The paper[1] represents equal contributions in terms of writing, but I provide additional discussion on how HitL participants interacted with the model.

10.1 Research objective

The current study was motivated by the findings in *Do Prosody Transfer Models Transfer Prosody?* (Sigurgeirsson and King, 2023) presented in Chapter 5. The results reported there suggest that typical PT models do not learn strictly transferable representations of prosody. Because of how these models are trained, where the reference matches the target utterance, they fail to generalise to the typical use cases: the reference employed may differ from the target in terms of either speaker (*cross-speaker*[2]) or text (**cross-text**). Because of this failure to separate prosody from other reference features, a PT model may produce a prosodic rendition that does not apply to the target text. Ultimately, this leads to degraded naturalness under the cross-text condition (Sigurgeirsson and King, 2023).

[1] The paper was jointly written by its authors. In particular, my co-author devised the third research question and contributed substantially to Sections 10.2.2.1, 10.3, 10.4.3.1, and 10.4.4.1

[2] The term *different-speaker* was previously used throughout Part I to refer to the same thing.

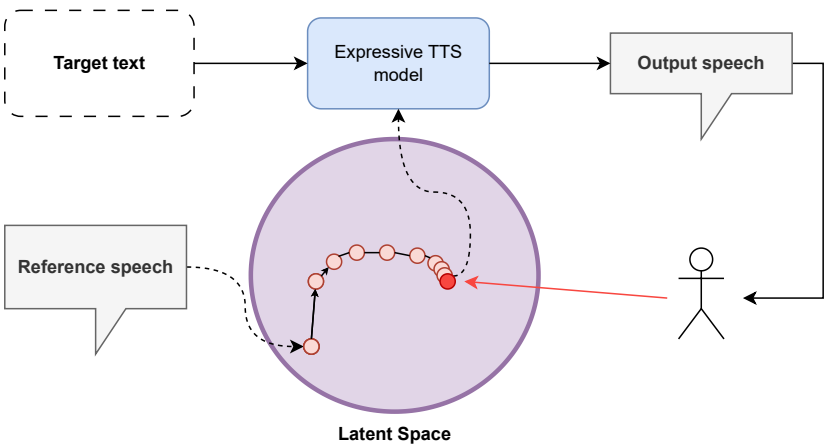


Figure 10.1: An overview of the prior work on employing a HitL for refining representations for expressive Text-To-Speech (TTS) (van Rijn et al., 2021). The refinement process is performed iteratively within the latent space until an optimal representation is found (red).

These issues raise questions about whether cross-text prosody transfer can ever be achieved *faithfully*. If the latent reference space models transferable representations of prosody, then a prosody representation that fits that goal should exist. But if it is not retrievable by conditioning on the reference prosody, it becomes a question about how else that representation can be retrieved from the reference space. One possible approach would be to directly incorporate human perception, utilising feedback from a HitL participant. In the current work, we investigated whether the prosody representations modelled by a representative PT model, **Daft-Exprt** (Zaïdi et al., 2022), can be improved by leveraging human perception, specifically for cross-text PT. We asked:

> **RQ 10-1:** *Can human-in-the-loop participants improve the perceived quality of cross-text prosody transfer?*

We considered a HitL-based method similar to the one proposed in van Rijn et al. (2021), where a HitL participant is tasked with adjusting emotive speech representations for improving emotive TTS. Their proposed method uses Gibbs sampling with human participants (GSP) (Harrison et al., 2020). Figure 10.1 illustrates how a GSP-based approach would be carried out for the PT task. An initial reference embedding is generated, and a HitL participant listens to the predicted prosodic rendition of the target text. The participant then makes iterative adjustments to isolated dimensions of the predicted reference embedding. At each iteration, the participant is asked to choose a

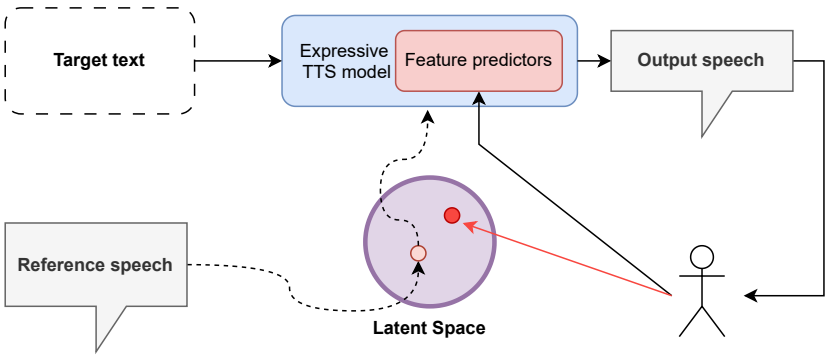


Figure 10.2: An overview of how we propose to employ HitL feedback for refining modelled representations. Instead of slow iteration within the latent space, we propose making adjustments in terms of known acoustic features (F_0, energy and duration). We expected prosodically similar utterances to be close in the latent reference space. The *edited* utterance (red) can be re-embedded in the latent reference space, which enables us to evaluate whether that was true for cross-text PT samples.

value for the given dimension, such that the rendition becomes prosodically similar to the reference. In van Rijn et al. (2021), participants cycle twenty times over all control features. Naturally, this approach is very time-consuming as previously discussed in Section 8.2.

Instead, we proposed grounding the participants' suggestions in terms of known acoustic features. In Harrison et al. (2020), a similar approach is proposed for emotional prosody generation, where the stimulus space is defined in terms of seven parametric manipulations of known acoustic features. That method is still based on Gibbs Sampling with People (GSP), requiring time-consuming iterative and isolated adjustments of the initially predicted rendition. Instead, we let HitL participants directly control a TTS model that separately models phone-level F_0, energy and duration. These are salient acoustic correlates of prosody, and together they cover the *control space* that our HitL participants interact with. This process is illustrated in Figure 10.2. We believed that HitL participants would be able to make appropriate adjustments to the initially-predicted cross-text PT rendition, and therefore improve perceived naturalness:

> **H10-1a:** *Human-in-the-loop participants can improve the perceived appropriateness of cross-text prosody transfer samples.*

> **H10-1b:** *Human-in-the-loop participants can improve the perceived naturalness of cross-text prosody transfer samples.*

However, participants might find that substantial changes are required to make the rendition appropriate for the target text. Such changes might make the suggested prosodic rendition less similar to the reference prosody, which would go against the goal of cross-text PT. Therefore, we asked:

> **RQ 10-2:** *Can HitL participants maintain the prosodic similarity with the reference?*

We hypothesised that HitL participants would be able to sense the prosodic function employed in the reference — in a way that a reference encoder could not — and make the adjustments while preserving their perception of this function in the output:

> **H10-2a:** *Human-in-the-loop participants can preserve the prosodic similarity with the reference while improving the perceived appropriateness of the prosodic rendition*

The control scheme used in the current study is based on the one proposed in Chapter 9. There, an Large Language Model (LLM) is instructed to make the adjustments instead of HitL participants. In the current work, we were interested in how the participants would fare using such a control scheme:

> **RQ 10-3:** *Is the suggested control scheme conducive for HitL interaction?*

We evaluated the participant effort involved in achieving the task using the proposed control scheme.

10.2 Proposed method

10.2.1 Baseline model architecture

The proposed HitL method does not require online supervision and can, therefore, be implemented using any pre-trained PT model that supports AFC control. We therefore chose ***Daft-Exprt*** (Zaïdi et al., 2022) as our baseline PT model, which was extended to enable the proposed HitL-based approach. We used HiFi-GAN (Kong et al., 2020) to convert the predicted mel-spectrogram to a waveform.

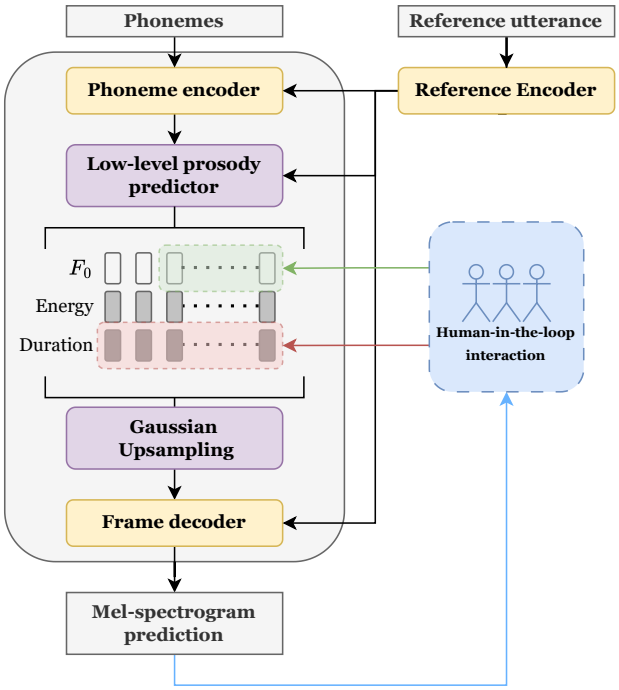


Figure 10.3: Overview of how information from the reference and adjustments by HitL participants affect model inference. Like in the original work Zaïdi et al. (2022), the reference embedding conditions speech generation in three locations (the phoneme encoder, the prosody predictor, and the frame decoder). HitL participants are asked to adjust word- and utterance-level F_0, energy, and duration features to improve the rendition for the cross-text PT task. Word-level adjustments are shown in green and utterance-level ones are shown in red. After each adjustment, participants can listen to the edited rendition and make further adjustments if required.

10.2.2 Human-in-the-loop approach

Given a reference and a *different* target text, an initial cross-text PT sample is generated. HitL participants interacted with the TTS model through the low-level prosody predictor, which predicts phone-level log-F_0, log-energy and duration. Manipulating individual phone-level predictions is believed to be too complex for the proposed HitL task (Mohan et al., 2021). So instead, we adopted a modified version of the control scheme discussed in Chapter 9, which enables both utterance- and word-level adjustments to the predicted acoustic feature values. The adjusted values are then provided to the acoustic model to complete synthesis.

10.2.2.1 Word-level control

Word-level control inputs were mapped to corresponding phone-level adjustments for compatibility with the modelling resolution of **Daft-Exprt**. We treated log-F_0 and log-energy adjustments in the same way, while treating duration adjustments slightly differently. First, the acoustic model predicts all phone-level log-F_0 and energy values:

$$v_1, v_2, \ldots, v_n$$

where v represents either phone-level log-F_0 or log-energy predictions henceforth. We then computed a word-level feature mean, K_w, and a per-phone scaling factor, s_i, for each input word $w = p_1, p_2, ..., p_n$:

$$K_w = \frac{1}{n}\sum v_i, \quad s_i = \frac{K_w}{v_i}, \quad i \in 1, \ldots, n \tag{10.1}$$

The HitL participant can then suggest changing the initial mean feature value K_w to a new one, K'_w. This results in per-phone changes such that:

$$v'_i = \frac{K'_w}{s_i}, \quad i \in 1, \ldots, n$$

This form of control, therefore, resulted in equal adjustments to all phones p_i in the word, proportional to their originally predicted value v_i. Voiceless phones were not included for the computation of K_w. Log-F_0 and log-energy values far outside the known training distribution can result in distorted output. We therefore determined a suitable range for both modifications based on the training corpus statistics. Before training, the training speaker's phone-level means and standard deviations (σ) for log-F_0, log-energy, and durations were computed. We found that limiting the F_0 range to $\pm 3\sigma$ and energy to $\pm 1.5\sigma$, relative to the training speaker's statistics, resulted in a good control range for this task. F_0 and energy control inputs were therefore limited such that no edited phone-level log-F_0 or log-energy value falls outside this range.

This control mechanism is counterintuitive for duration, since it would set the initial word duration to its phone-duration mean. Therefore, we resorted to a straightforward approach for duration control. Duration of each phone in the word by a constant within the range $[0, 2]$[3]. That is, each word could be made up to twice as long in duration.

[3]This design technically allowed participants to set the duration of any word to 0 s, but no HitL participant actually did.

10.2.2.2 Utterance-level control

Like in (Sigurgeirsson and King, 2024), we also provided utterance-level controls for log-F_0, log-energy, and duration. We believed that applying global changes to prosody may help participants make fast adjustments relevant to conveying emotions, expressions, and speaking styles. Again, we treated log-F_0 and log-energy control in the same way. Utterance-level control inputs were applied through word-level adjustments: a participant-submitted utterance-level control input was converted into each word's corresponding word-level control inputs. We determined the utterance-level control range such that any resulting word-level change remains within the statistical per-phone ranges described in Section 10.2.2.1. The utterance-level duration control simply allowed participants to scale all phone durations equally within the range $[0, 2]$.

10.3 Experimental setup

10.3.1 Baseline model training

The baseline **Daft-Exprt** model was trained on the 2013 Blizzard Challenge corpus (King and Karaiskos, 2013), an expressive audiobook corpus. The corpus is single-speaker, comprising over 300 hours from a professional English female voice actress. We used only the segmented split of the corpus, which includes 52 hours from 55 different books. Utterances shorter than 0.3 seconds and longer than 15 seconds were excluded. The remaining approximately 40,000 utterances, comprising around 40 hours, were used to train the baseline PT model. We followed the original **Daft-Exprt** preprocessing procedure (Zaïdi et al., 2022). Alignments were estimated using the Montreal forced aligner (McAuliffe et al., 2017), phone-level log-F_0 was estimated using REAPER[4], and the l^2-norm of spectrogram frames was extracted for energy.

The model was trained for 24 hours, distributed over 4 NVIDIA Tesla V100-SXM2-16GB GPUs, using a batch size of 96 samples. We trained a HiFi-GAN Kong et al. (2020) vocoder using ground-truth utterances from the same corpus used for training the acoustic model. The vocoder is trained for 12 hours on a single NVIDIA Tesla V100-SXM2-16GB GPU using a batch size of 8 samples.

[4]https://github.com/google/REAPER

10.3.2 Collection of Human-in-the-loop edited samples

In the proposed method, HitL participants first listened to the reference utterance used to generate the cross-text PT sample. They were then asked to suggest adjustments to make the resulting prosody more appropriate for the target text — while preserving the perceived prosodic similarity with the reference. Taking inspiration from existing User Interfaces (UIs) for TTS system interaction (Kondo and Morise, 2019; Tits et al., 2021), we developed a web-based UI using Streamlit[5], to facilitate the proposed HitL method. Control of both utterance- and word-level features is realised using *slider* UI elements. Control inputs received through the UI are used to compute the corresponding log-F_0, log-energy and duration values, as explained in Sections 10.2.2.1-10.2.2.2.

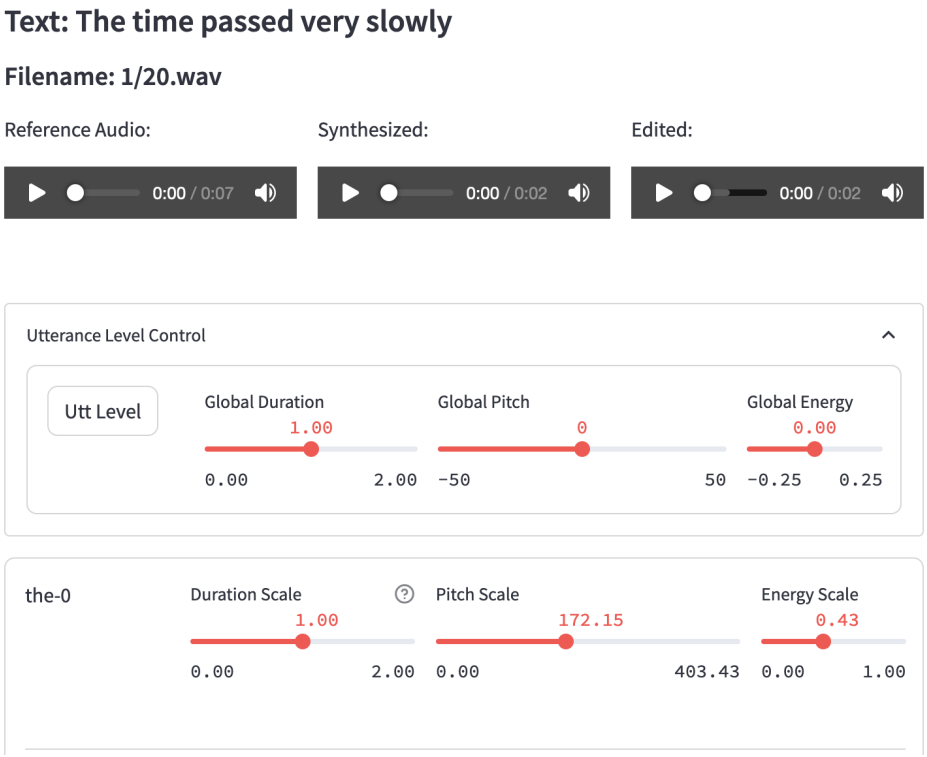


Figure 10.4: A screenshot from the UI employed for the HitL-based task, showing utterance-level control elements and controls for the first word in the target text. Control of all features is achieved through slider elements, limited to a pre-defined range. After any slider value is changed, a new rendition is synthesised based on the adjusted feature value.

To account for any possible individual bias, we asked all participants to adjust the same list of stimuli. We chose five reference utterances that would each yield a perceptually-distinct prosodic rendition. We used four target texts that would typically elicit a par-

[5]https://streamlit.io/

ticular prosodic effect. All chosen target texts are short (5 words or fewer) to keep the HitL process as simple as possible and maximise the number of collected samples. The five references and four target texts correspond to $5 \times 4 = 20$ reference-target pairs. These were employed to generate the 20 cross-text PT samples used in this experiment. More information about the stimuli used in our experiments is given in Appendix D.

After making adjustments, the resulting version could be played back, as many times as the HitL-participant deems necessary, to determine whether the suggested adjustments improved the rendition. Once the participant determined that no further| improvements could be made, the **edited** sample and the initially synthesised cross-text PT sample (**original**) were saved. HitL participants were asked whether they believed they could make appropriate adjustments using the control scheme in the first place. We also asked participants to indicate how confident they were (*"low"* or *"high"*) in their suggestions. We recruited 33 HitL participants, with a mixed language background[6], to participate in the experiment. Their participation resulted in $5 \times 4 \times 33 = 660$ **original / edited** sample pairs.

10.3.3 Evaluation of edited samples

To test **H10-1a**, we measured the perceived *appropriateness* of both the **original** and **edited** prosodic renditions. We employed an A/B preference design where raters were asked to base their preference on how appropriate the prosodic rendition is for the synthesised text. We also evaluated perceived naturalness — of both **original** and **edited** samples — in a standard Mean Opinion Score (MOS) listening test, to test for a positive correlation between prosodic appropriateness and naturalness (**H10-1b**).

We anticipated that raters could make these improvements on the cross-text PT samples while preserving prosodic similarity with the reference (**H10-2a**). We conducted a listening test based on a Multi Stimulus with Hidden Reference and Anchor (MUSHRA) design to test this claim. We included a PT sample that employed a random reference as a hidden anchor. Raters were asked to first listen to the reference utterance, used to generate the **original** cross-text PT sample, before evaluating the **original** sample, the **edited** sample, and the random anchor. Raters are asked to indicate, on a scale from 0 to 100, how prosodically similar these samples are to the reference. They are asked to base their judgement on the same qualities used in Skerry-Ryan et al. (2018): (1) the

[6] An indeliberate lack of control.

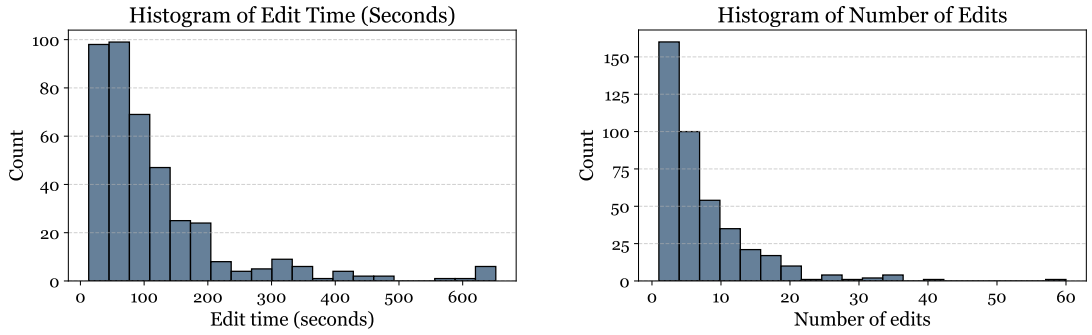
pitch, and the intonation throughout the utterance; (2) stress, both word and syllable stress reflected either in loudness or change in pitch; (3) speaking rate, and how it may change over time; and (4) the length of pauses.

Of the 660 **original/edited** pairs, 193 ($\approx 30\%$) were unmodified by the HitL participant. In these cases, the HitL participants deemed the **original** prosodic rendition good enough. These unmodified samples, and 47 additional unusable pairs, were removed from the set before collecting ratings for preference, prosodic similarity, and naturalness from a second group of independent raters. This resulted in 840 MOS questions, 420 MUSHRA-like screens, and 420 A/B preference questions. We recruited 68 native UK/US raters via Prolific[7] to participate in the study. Each sample is evaluated by at least three raters and at most five.

10.4 Results

10.4.1 RQ 10-3: Is the suggested control scheme conducive for HitL interaction?

We analysed how participants interacted with the model and how much effort was required to achieve a suitable rendition. We also asked participants to self-report their confidence in the **edited** rendition being better than the **original** one.



(a) The time used by HitL participants to complete each trial.

(b) The number of edits, per trial, made by HitL participants.

Figure 10.5: Histograms of time spent by HitL participants and the numbers of edits they made for each cross-text PT trial.

Participants spent, on average, 123.5 (SD : 138.0) seconds modifying each sample, performing 7.0 (SD : 7.1) individual operations per utterance on average to complete

[7] https://www.prolific.com/

the task. The histograms shown in Figure 10.5 indicate that the effort required to make the adjustments varied substantially between HitL participants.

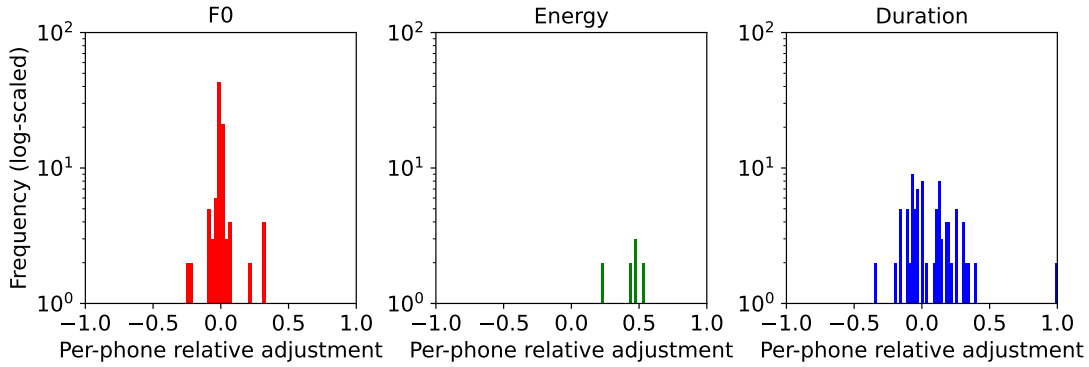


Figure 10.6: The distributions of phone-level control-inputs for F_0, energy and duration (unchanged values omitted). These results suggest that energy control was substantially less important to HitL participants than F_0 and duration control.

As previously explained, 193 cross-text samples were considered good enough by HitL participants and removed from the set. Of the 420 filtered pairs, participants indicated a *high confidence* in their performance for 82.6% and a *low confidence* for 17.4%. Interestingly, as visualised in Figure 10.6, participants made approximately three times fewer edits to predicted energy values than F_0 and duration values. The lack of interaction could indicate that HitL participants did not find energy to be a salient control feature for completing the task.

10.4.2 RQ 10-1: Can human-in-the-loop participants improve the perceived quality of cross-text prosody transfer?

10.4.2.1 Perceived appropriateness (H10-1a)

Edited samples were expected to improve perceived appropriateness when compared to **original** ones (H10-1a). The A/B preference results, and other results from the follow-up perceptual evaluations, are reported in Table 10.1.

To account for `rater` variation and repeated measures, we performed an analysis based on Highest Density Intervals (HDIs). A Bayesian mixed effect model, with random intercepts for `rater` and `utteranceID`, was trained to estimate `system` (**original** or **edited**) preference probabilities across the evaluation set. We break the results down by reported participant confidence.

Overall, including both *low* and *high confidence* samples, listeners indicated a preference for the **edited** samples compared with the **original** ones, 57.9% vs. 42.1%. This is a marginal but credible difference (95% HDI: $[0.05, 0.44]$), as the highest density interval does not include 0. For *low confidence* samples, the preference rates for **edited** and **unedited** samples converged, and the two are not credibly different (95% HDI: $[-0.22, 0.18]$). But when HitL participants indicated a high confidence, the preference for **edited** samples is particularly prominent. Raters credibly preferred **edited** samples at a rate of 59.6% to 40.4% of **original** samples under this condition (95% HDI: $[0.17, 0.57]$).

In summary, participants could improve the perceived appropriateness of cross-text PT samples. This effect was especially clear when HitL participants indicated high confidence in their suggestions:

> **H10-1a**: *Human-in-the-loop participants can improve the perceived appropriateness of cross-text prosody transfer samples.* **Accept** ✔

Table 10.1: Main subjective results broken down by HitL participant confidence. The highlighted row is for the *high confidence* results of the proposed method.

Confidence	Condition	MOS	A/B preference	PT-MUSHRA
Low confidence	**original**	-	50.7%	60.6 ± 3.5
	edited	2.7 ± 0.2	49.3%	52.4 ± 4.3
	random	-	-	37.7 ± 4.8
High confidence	**original**	-	40.4%	58.1 ± 1.4
	edited	$\mathbf{3.0 \pm 0.1}$	**59.6%**	$\mathbf{55.3 \pm 1.5}$
	random	-	-	35.2 ± 1.7
Overall	**original**	3.2 ± 0.1	42.1%	58.6 ± 1.3
	edited	3.0 ± 0.1	57.9%	54.0 ± 1.5
	random	-	-	35.6 ± 1.7

10.4.2.2 Perceived naturalness (H10-1b)

We hypothesised that increased perceived appropriateness would result in increased naturalness (**H10-1b**). To evaluate differences in mean MOS between `systems` (**original** or **edited**), we fitted a Linear Mixed-Effects (LME) to the MOS scores, including `system` as a fixed effect, with random intercepts for `rater` and `utteranceID`.

We find that, overall, **original** samples are perceived marginally more natural than **edited** ones ($\beta = 0.22, \text{SE} = .043, p < .001$). We then isolated the comparison to

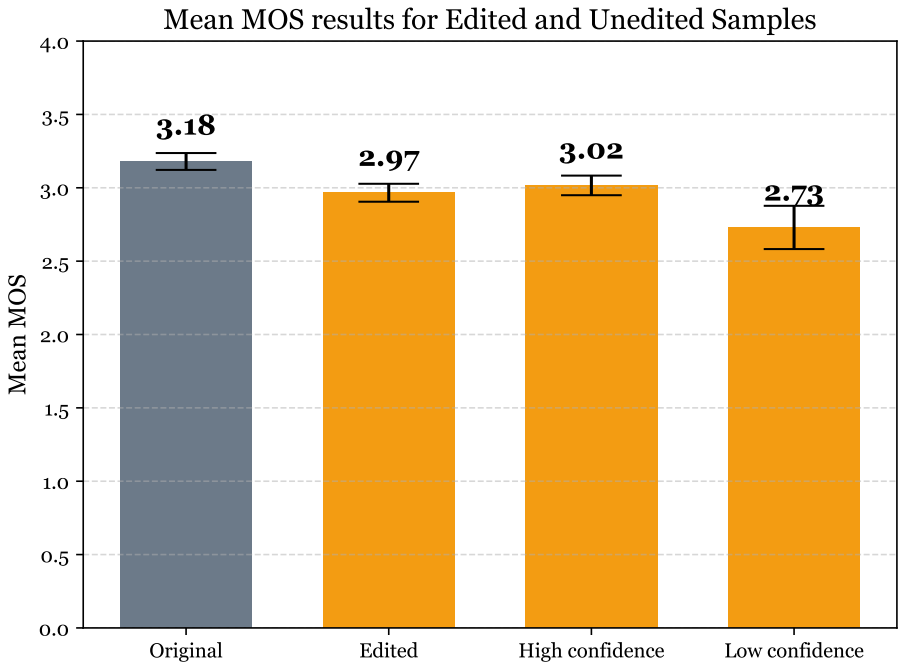


Figure 10.7: MOS scores **original** (blue) and **edited** samples (orange) samples, broken down by reported HitL participant confidence.

just utterances where the HitL participants had indicated high confidence. The corresponding **original** samples, which employed the same references, were used in this comparison. We did not observe an improvement in reported mean MOS under this isolation as the **edited** samples are still marginally less natural than **original** ones ($\beta = 0.16, \mathrm{SE} = .058, p < .001$). These results suggest that improving the renditions' appropriateness was insufficient to improve the cross-text samples' perceived naturalness:

> **H10-1b**: *Human-in-the-loop participants can improve the perceived naturalness of cross-text prosody transfer samples.* **Reject ✗**

During development, we noticed that the proposed control procedure occasionally resulted in artefacts and acoustic distortion, despite our attempts to limit the control range of acoustic features to those that reflect the training corpus statistics. It is not unlikely that this distortion biased listeners' naturalness judgements.

10.4.3 RQ 10-2: Can HitL participants maintain the prosodic similarity with the reference?

We tasked our HitL participants to appropriately adjust the **original** samples, while not diminishing the prosodic similarity with the reference. We hypothesised that the proposed control scheme would allow participants to do so (**H10-2a**). An LME model was fitted on the MUSHRA-like scores, with `system` (**original**, **edited**, and **random**) as a fixed effect. Random intercepts were included for `rater` and `questionID`. **Random** was used as the reference `system` and Likelihood Ratio Tests (LRTs) were conducted for any additional pair-wise comparisons. It should be noted that this design does not account for variation induced by any difference between HitL participants.

As expected, **random** samples were rated much less prosodically similar to the reference than either **original** samples ($\beta = 22.93, \text{SE} = .79, p < .001$) or **edited** samples ($\beta = 18.17, \text{SE} = .92, p < .001$). An LRT showed that **edited** samples were perceived marginally less prosodically similar to the reference, when compared to **original** ($\chi^2(1) = 28.68$, $p < .001$).

We then limited the analysis to just ratings where HitL participants indicated a high confidence in their suggestion. **Edited** samples were again rated slightly lower than **original** samples (55.3 ± 1.5 vs. 58.1 ± 1.4). But, a pairwise LRT showed that this difference is not significant ($\chi^2(1) = 0.91$, $p = .340$).

> **H10-2a**: *Human-in-the-loop participants can preserve the prosodic similarity with the reference while improving the perceived appropriateness of the prosodic rendition* **Accept ✔**

It should be emphasised that this only applied to *high confidence* samples, representing 82.6% of evaluated samples and 52.5% of all samples.

10.4.3.1 The divergence between latent and perceived similarity

The prosody transfer task is premised on the model's ability to construct a latent reference space, from which transferable representations of prosody can be sampled. Given this premise, we would expect prosodically similar utterances, based on similarity judgements, to yield similar reference embeddings. If this is not observed, it suggests that the reference encoder may be capturing features other than prosody, which may nonetheless correlate with prosody under certain conditions.

Based on the foundational PT premise, we hypothesised that samples rated as prosodi-

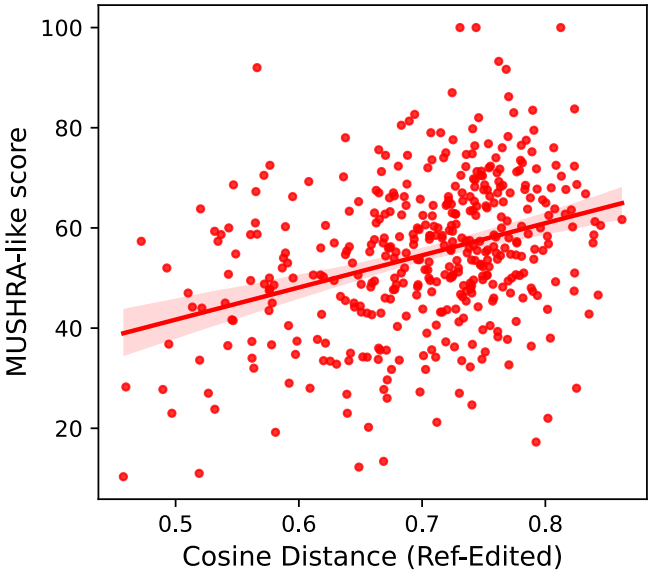


Figure 10.8: MUSHRA-like scores of edited cross-text PT samples plotted against the cosine distance between the original reference embedding and the embedded edited cross-text PT sample. A linear regression model is fitted to this data, and the shaded red area indicates the 95% confidence interval.

cally similar to the reference would be closer to the reference than those samples rated less similar. Interestingly, we observed the opposite, as illustrated in Figure 10.8. The figure shows the relationship between similarity judgements of a reference and a corresponding **edited** cross-text PT sample on the Y-axis, and measured cosine distance between the reference embedding and the embedding corresponding to the **edited** sample on the X-axis. There is a moderate correlation between the perceived prosodic similarity and embedding cosine distance. Therefore, **edited** samples that are perceived to be prosodically similar to the reference tend to be further away from it in reference space.

We hypothesised that these results demonstrate two things: (1) HitL-participants were able to identify a *"prosodic intent"* from the reference and *faithfully* modify the cross-text PT sample with regard to the identified intent and the target text; and (2) closeness to a reference embedding is not a reliable metric for prosodic similarity for cross-text PT.

10.4.4 Post-hoc analysis of user interaction

10.4.4.1 User Effort

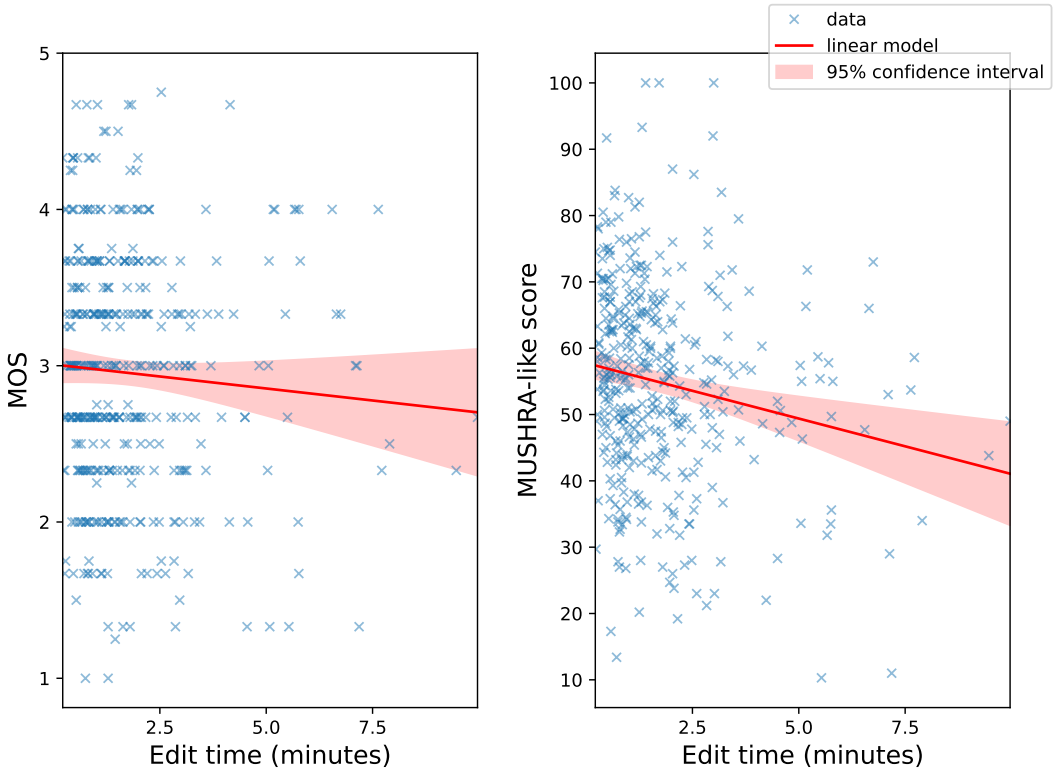


Figure 10.9: The relationship between HitL effort and the quality of the output. **Left**: the line of best fit for perceived naturalness as a function of HitL effort **Right**: the line of best fit for PT quality as a function of HitL per-trial duration.

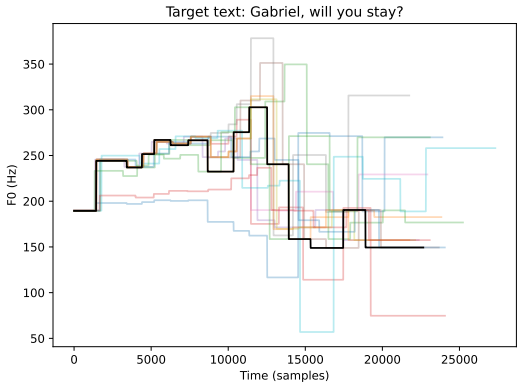
We also studied the relationship between HitL participant effort and the perceived quality of the output. Here, we took the time spent editing a sample as an indication of a participant's effort. We assumed that perceived naturalness would be positively correlated with the overall effort of HitL participants. However, this is not the case, as demonstrated by Figure 10.9. We fitted a linear regression model to the reported naturalness results as a function of the time taken to make the modifications. Contrary to our assumption, HitL effort is moderately negatively correlated with perceived naturalness. A comparable analysis of perceived prosodic similarity revealed a stronger negative correlation with participant effort. These results have many plausible reasons, but they underscore the importance of minimising user effort for HitL tasks for TTS.

Editor Tunnel Vision

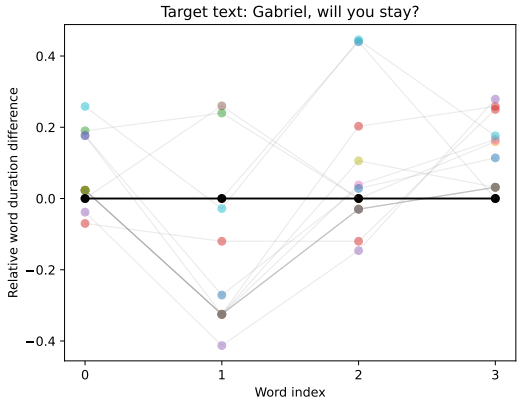
This negative correlation between effort and output quality could be explained by an *optimisation drift*: where some HitL participants grounded their $t+1$th edit on the rendition corresponding to the tth edit and so on, drifting further and further away from the original rendition. The drift may have resulted in the iterative degradation that the results suggest. Whether true or not, minimising required user effort is essential for this task.

10.4.4.2 User agreement

Later, I investigated how participants interacted with the control scheme. Different HitL participants suggested a wide range of different renditions for the same cross-text PT sample. An illustrative example of this is visualised in Figure 10.10, which shows how different HitL participants make F_0 and duration adjustments for the same sample. The samples submitted by participants vary substantially, both in terms of F_0 and duration modifications. The variation likely reflects the individuals' perception of the reference prosody and their interpretation of the target text.



(a) F_0 modifications: black contour represents the original F_0 word-level means (K_w in Section 10.2.2.1). Other contours are edits made by HitL participants.



(b) Duration modifications: the black horizontal line represents the initial word durations, whereas coloured circles represent the proposed word-duration modifications.

Figure 10.10: An example of how different HitL participants suggested different modifications to F_0 and duration.

The HitL participant *agreement*, in terms of the direction of the change, was objectively analysed. Here, participants agree if the sign of the difference between the original and suggested feature value is the same. For example, if two editors decide to increase the duration of a specific word, regardless of how much, they are in *agreement* on that word's duration. Word-level agreement is visualised in Figure 10.11. *Full* word-level

agreement occurs when all participants agree on the modification direction from the initially predicted value. Full word-level agreement corresponds to an agreement value of 1. A minimum agreement, corresponding to 0.5, occurs when the participant group is split exactly in half on how to change the word.

The results indicate less editor agreement regarding modifications to F_0 and duration than energy modifications. This reduction in editor agreement is interesting in light of the results in Section 10.4.1, which showed that HitL participants prefer making F_0 and duration modifications, rather than energy modifications, at an approximate rate of three to one. Taken together this may suggest that: (1) F_0 and duration are salient features for the task, while energy is not, (2) changes to salient features allow HitL participants to submit modifications based on their perception of the reference prosody and how it fits with the target text; and (3) difference in perception yields lower agreement rate for salient control features.

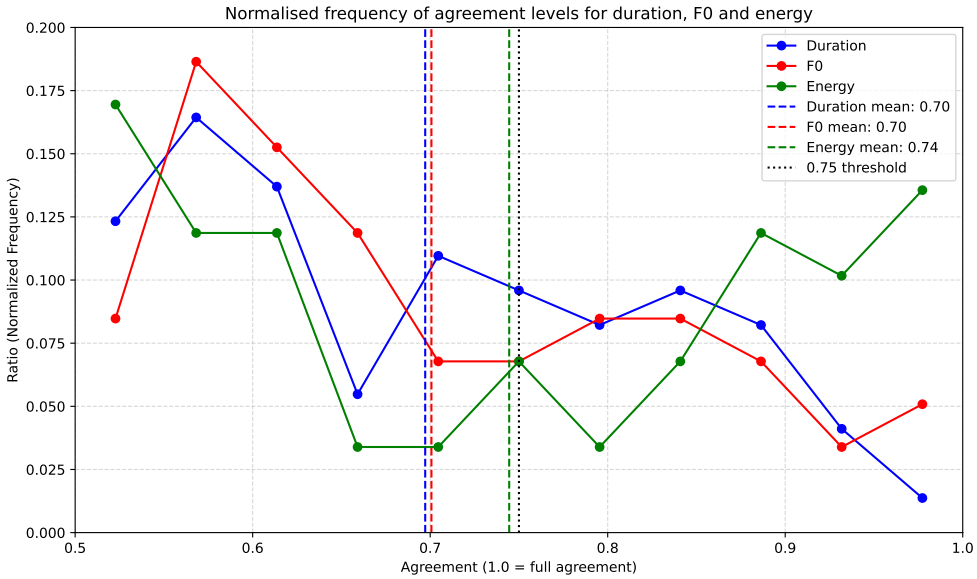


Figure 10.11: A line-plot representation of HitL participant agreement histograms. The histograms are agreement-normalised (bins sum to 1), such that a higher ratio (Y-axis) indicates a higher proportion of samples having the specified level of agreement (X-axis). Dashed lines indicate mean word-level agreement for each feature.

10.5 Conclusion

We initially asked: **RQ 10-1**: ***Can human-in-the-loop participants improve the perceived quality of cross-text prosody transfer?*** The results show that participants can make the output prosody more appropriate for the target text. However, this does not

lead to an improvement in perceived naturalness. We believe the unnatural artefacts introduced through our control mechanism, as mentioned in Section 10.4.2.2, may have contributed to this negative result. We also asked: **RQ 10-2:** ***Can HitL participants maintain the prosodic similarity with the reference?***. The results indicate that the HitL participants managed to make the adjustments without diminishing the perceived prosodic similarity of the output with the reference.

Lastly, we explored **RQ 10-3:** ***Is the suggested control scheme conducive for HitL interaction?***. Results in Section 10.4.1 indicate that F_0 and duration are more salient control features for this task than energy. Later, in Section 10.4.4.2, I showed that these salient features are connected with a lower level of editor agreement than less salient ones. I hypothesise that this reflects the editors' perception of the reference prosody and how to apply it to the target text. Results shown in Figure 10.9 suggest that more interaction with the model reduces naturalness, highlighting the importance of minimizing user effort for effective HitL applications in TTS. Despite limitations in the control scheme, the findings support its potential for use in HitL scenarios.

Chapter 11

Discussion

The two studies discussed in the current part employed an Acoustic Feature Control (AFC)-model to improve listeners' perception across various tasks. Models like **FastSpeech 2** allow for nuanced control of acoustic correlates of prosody. However, the proposed control scheme, used in both studies, makes several simplifications. For example, the scheme does not enable modification of pause durations or changing word-level F_0 contours in a sophisticated manner. Therefore, the control scheme only allows for generating a subset of the full range of plausible prosodic renditions for a given target text. The proposed control scheme was kept simple to make the AFC-task easier to accomplish. Results presented in Section 10.4 emphasise the importance of this principle, as Human-in-the-Loop (HitL) effort is shown to negatively correlate with perceptual objectives.

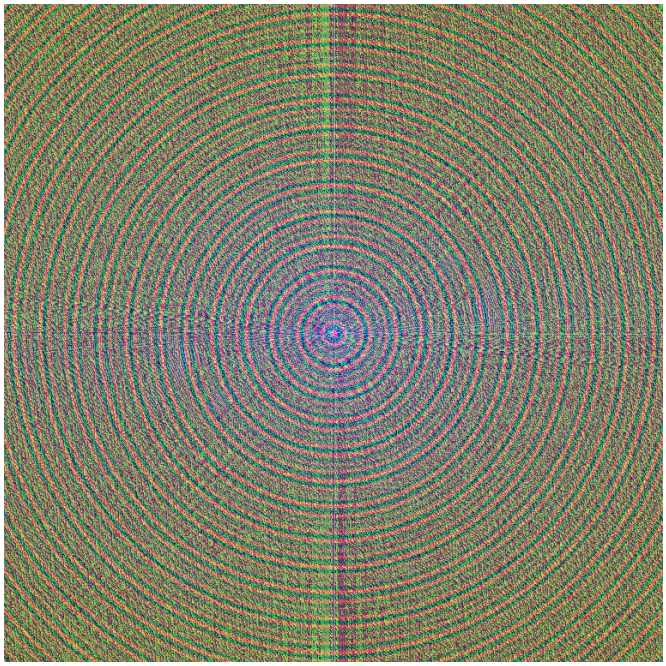
The results demonstrate that there is, despite the simplifications employed, substantial effort required to control Text-To-Speech (TTS) models in this manner. HitL participants required, on average, approximately two minutes to complete the task in Chapter 10. So, AFC models may not be suitable for applications that require a responsive control method. But AFC-control can be converted into simpler forms of prompt-based control like the one demonstrated in Chapter 9.

Despite these limitations, the control scheme demonstrates improvements for several tasks: (1) cross-text Prosody Transfer (PT), (2) expressive conversational speech, and (3) speaking style generation. Results in Sections 9.6.2-9.6.3 indicate that this improvement is particularly salient when contextual information or propositional content calls for vocal renditions uncommon in the training data. There are, of course, limits to

how far away from the training statistics an AFC model can be pushed. Conditioning an AFC model on feature values atypical of the training corpus can lead to degraded naturalness, as discussed in Section 10.4.2.2. Despite this, the results from both studies indicate that the control scheme allows for producing meaningful prosodic variation — within the range of the training statistics — that listeners perceive as more appropriate than the model-predicted variation.

Part III

Prompt-Based Control



We have studied and examined reference-based models in Part I and Acoustic Feature Control (AFC) models in Part II. Both of these control strategies have been the subject of extensive research. In this part, we turn to the more recent development of *prompt-based* control for TTS, which seeks to guide the TTS model through a natural language description of the desired output.

Published papers presented in this part:

Chapter 13: *RepeaTTS: towards feature discovery through repeated fine-tuning* (Sigurgeirsson and King, 2025)

Chapter 12

Introduction & background

12.1 Introduction

Part II introduced a control strategy for expressive Text-To-Speech (TTS) based on the direct control of known acoustic features. While such strategies are highly interpretable and precise, experimental results revealed practical limitations. Acoustic Feature Control (AFC) tasks are technically complex, and users frequently struggle to control the rendition using this type of control successfully (see Chapter 10). Even if participants clearly understand what an appropriate vocal rendition might sound like, based on some contextual information, articulating that knowledge through acoustic feature parameters is unintuitive.

Large Language Models (LLMs) (e.g., Devlin et al., 2019; Brown et al., 2020; Ouyang et al., 2022) have demonstrated strong few-shot learning capabilities across a wide range of language tasks when scaled to sufficient size (Brown et al., 2020). Their flexibility comes from the ability to solve tasks through *prompting*: providing a task description and input directly in natural language, without any fine-tuning or parameter updates. In this approach, the task description and the task itself (collectively known as the 'prompt') are used to query the LLM for a solution. LLMs have been used to solve tasks as diverse as text completion, answering factual questions, translation, and grammar correction (Brown et al., 2020), among others. Despite the versatility of LLMs, they operate on the same principle as smaller language models: given an input text, they predict a likely output text continuation. Thus, prompting simply refers to supplying a LLM with a query (the input text), from which the LLM predicts a likely answer (the output text).

Prompting of LLMs offers an intuitive control method: the target behaviour can be specified through just natural language. Therefore, many different fields of work have aimed to provide this form of *prompt-based* control for modalities other than text, including for *prompt-to-image* (Ramesh et al., 2021) and *prompt-to-music* (Agostinelli et al., 2023). Several recent studies (e.g., Guo et al., 2023; Yang et al., 2024; Lyth and King, 2024; Lacombe et al., 2024) explore prompt-based control in TTS, enabling users to influence characteristics of the generated speech — such as the perceived gender or the speaker's pitch — through natural language descriptions.

Prompt-based TTS may offer a user-friendly approach. But like the other TTS control strategies discussed in this thesis, prompt-based control is constrained to features on which the model has been trained. Yet, unlike typical reference-based and AFC models, many recent prompt-based TTS models can generate a range of plausible prosodic variations for the same control inputs; variations shaped by corpus statistics that are beyond the user's control.

This part begins with a brief overview of prompt-based TTS in Section 12.2. Chapter 13 then explores how such uncontrolled variation can be discovered and incorporated into the model through repeated fine-tuning.

12.2 Background

12.2.1 Prior influence of LLMs in TTS

Even before prompt-based TTS models, many sought to leverage the representational power of LLMs to improve TTS. LLMs (e.g., Devlin et al., 2019; Brown et al., 2020; Ouyang et al., 2022), utilise Contextual Word Embedding (CWE) for word or token representations. These CWEs encode semantic and syntactic word relations within a context (input text). CWEs have been used for improving prosody prediction generally (Karlapati et al., 2022; Xiao et al., 2020; Chen et al., 2021; Makarov et al., 2022), and in more targeted tasks such as style or emotion modelling (Prateek et al., 2019; Yoon et al., 2022), or dialogue TTS (Guo et al., 2021). While CWEs offer richer text representations than context-independent embeddings, they remain fixed for a given input text and cannot, by themselves, support varied renditions of the same text. Prompt-based TTS achieves this by conditioning speech generation on different natural language descriptions of the output.

12.2.2 Types of control

Prompt-based models aim to control some form of non-lexical variation based on *free-form* descriptions of the variation. That is, the prompt describes aspects which are not determined by the text itself. Which non-lexical aspects (henceforth *control features*) are targets for control differ between different prompt-based TTS models.

PromptTTS (Guo et al., 2023) provides utterance-level control of perceived gender, pitch, speaking rate, loudness, and emotion. InstructTTS (Yang et al., 2024) focuses on emotional content and speaking style. Lyth and King (2024) provides control of the speaker's accent, pitch, speaking rate, and recording quality. **ParlerTTS** (Lacombe et al., 2024), a model based on that of Lyth and King (2024), aims to additionally control perceived speaker identity. Speaker identity control is achieved by using a speaker-labelled training corpus and then including speaker labels in the description prompt. Although most prompt-based TTS models offer an utterance-level form of control, where the description affects suprasegmental features of the utterance as a whole, some approaches aim to provide more temporary fine-grained manipulation of speech. But fine-grained control is often limited to just a single control feature, like lexical stress (e.g., Jin et al., 2024; Zhou et al., 2024).

To learn the *prompt-to-acoustics* mapping, prompt-based models are invariably trained on a *prompt-annotated* speech corpus. That is, each training utterance is labelled with a natural language description of the settings of the control features. This is done to support *free-form* natural language control; training a prompt-based model on templated descriptions of the control features (e.g., *loudness: 3, pitch: high, speaking rate: low*) does not generalise to diverse, free-form descriptions. Such description-annotated corpora are not common, so much work on prompt-based TTS involves annotating an existing speech corpus without such descriptions.

InstructTTS (Yang et al., 2024) uses human-generated annotations of emotional content and speaking styles. PromptTTS (Guo et al., 2023) employs a similar strategy, collecting expert-written free-form descriptions of the speaker, manner of speech, and emotional content. Collecting these annotations is a time-consuming process, so human-annotated speech corpora are typically limited in size (44 hours in Yang et al. (2024)). Because of this, and the underlying lack of existing prompt-labelled speech corpora, many models are trained on LLM-generated descriptions instead (e.g., Leng et al., 2024; Ji et al., 2024; Lyth and King, 2024; Seshadri et al., 2022). First, a limited

set of control features is extracted from training utterances using external methods. Then, for each training utterance, an LLM is prompted with (1) the task description, which explains to the LLM what sort of descriptions it should generate, and (2) the setting of the control features for the current training utterance. This more time-effective approach allows for annotating much larger speech corpora; for example, Lyth and King (2024) annotate over 45,000 hours of speech. At the same time, this approach leverages the ability of LLMs to generate diverse but coherent text continuations for the same LLM text input (task description and control feature settings) to support free-form control; the LLM predicts descriptions of the same feature settings, phrased in different ways.

12.2.3 Model architecture

Like any other conditioning model (e.g. reference-based models in Part I), prompt-based TTS models require a strategy for conditioning speech generation on the user input. There are various approaches to implementing description prompt conditioning, and the selected method is typically influenced by the underlying TTS architecture. PromptTTS (Guo et al., 2023) is an encoder-decoder TTS architecture. There, the description prompt is separately encoded into a fixed-size embedding and prepended to the latent phoneme sequence before speech decoding. PromptStyle (Liu et al., 2023), a flow-based TTS architecture, learns a *style embedding space* from reference speech. It then trains a prompt encoder that is aligned with the style embedding space using a cosine similarity loss. InstructTTS (Yang et al., 2024), a diffusion-based (Ho et al., 2020) TTS architecture, predicts scaling and biasing parameters from an instruction prompt. These parameters are used to condition both phoneme encoding and mel-spectrogram diffusion, using a strategy similar to FiLM conditioning (Perez et al., 2018).

Many prompt-based TTS models are based on a decoder-only audio codec language model architecture (Yang et al., 2024; Lyth and King, 2024; Seshadri et al., 2022), sometimes referred to as Speech Language Models (SLMs). These models predict audio *codes* in an autoregressive manner, which formulates the speech generation task similarly to language modelling. SLMs are not limited to prompt-based models; instead, they form a general framework for TTS which can use alternative inputs for conditioning, such as voice prompts (Wang et al., 2023). The baseline model architecture used in the upcoming Chapter 13 is of this type, and is based on the architecture

proposed by Lyth and King (2024). Due to its relevance, this particular SLM architecture is explained in detail below.

12.2.3.1 Lyth & King (2024)

Lyth and King (2024) proposes a SLM architecture which predicts speech for a target text, enabling control of several features based on a user-submitted natural language description. Like LLMs, SLMs operate on discrete feature representations. LLMs are trained on tokenised inputs, which are discrete representations of text units. Given a sequence of input tokens, an LLM predicts a likely next token. To formulate the speech generation task in the same way, speech inputs must be tokenised. Lyth and King (2024) use the Descript Audio Token (DAC) (Kumar et al., 2023) to discretise speech into discrete *speech tokens* at a frame rate of 86 Hz. DAC employs Residual Vector Quantisation (RVQ), where each speech segment is encoded into a token comprising 9 different codewords (fixed-size embeddings) selected from 9 codebooks. During decoding, the RVQ-based decoder requires all 9 codewords to reconstruct the speech segment at a given time step: the kth codeword is generated from the residual of the previous $k-1$th codeword. To account for this causality, Lyth and King (2024) employ a *delay pattern* (Copet et al., 2023), ensuring that the prediction for the kth codebook at time step t is conditionally dependent on the $k-1$th codebook at the same time step.

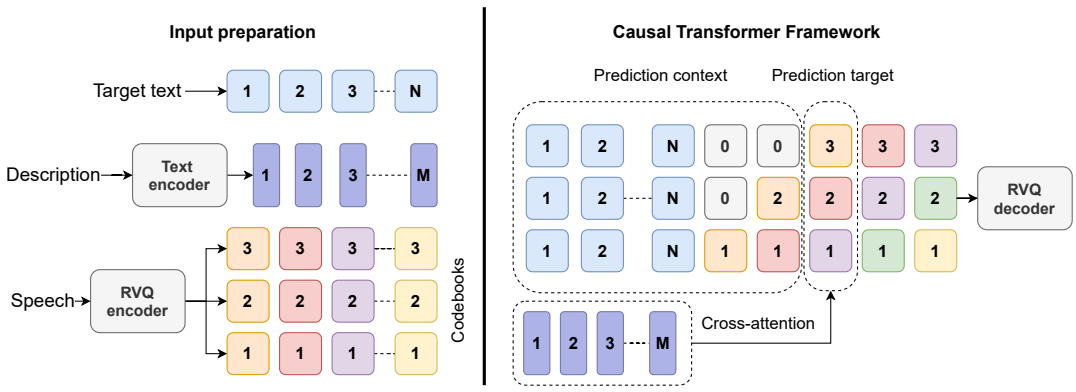


Figure 12.1: An overview of how inputs are prepared (left) and how the speech modelling task (right) is set up in Lyth and King (2024). First, the target text is tokenised into N tokens, the description prompt is encoded as a latent sequence of length M, and the target speech is tokenised using an RVQ-encoder (3 codebooks in this example, different time steps represented by different colours). Lyth and King (2024) employ a delay pattern to make the prediction of residual codewords conditional on higher-level ones. The prediction at every timestep is conditional on the description prompt representation, utilising cross-attention. When RVQ codewords have been predicted, they are reorganised before speech is decoded using an RVQ-decoder.

In Lyth and King (2024), a causal Transformer (Subakan et al., 2021) network is trained to predict the RVQ token indices in an autoregressive manner. To enable TTS, a tokenisation of the corresponding transcript is prepended to the speech token sequence of training utterances. This accounts for the causal nature of the transcription while allowing the prediction of each speech token to attend to every transcription token. Prompt conditioning is achieved using cross-attention instead. The description prompt is first tokenised and encoded using a frozen T5 text encoder (Chung et al., 2024). Then, the prediction of each speech token attends to the encoded description prompt. After predicting all codewords, they are reorganised into their corresponding token before being decoded into speech using an RVQ-based decoder.

The model is trained on over 45,000 hours of speech. Before training, a host of features were extracted for each training utterance: Signal-to-Noise Ratio (SNR), reverberation estimation, PESQ scores Rix et al. (2001), SI-SDR Le Roux et al. (2019) and pitch estimation using the PENN library Morrison et al. (2023). Using corpus metadata and the extracted features, training utterances were next labelled with descriptions of the speaker's accent, pitch, and speaking rate, as well as recording quality. Continuous features, such as speaking rate, are categorised into a small number of bins based on the training corpus statistics. Categorical features, such as the speaker's accent, do not require any processing; they can be included directly in the prompt. Finally, an LLM is employed to generate free-form descriptions of each utterance's setting of the control features.

They demonstrate that the model can generate highly natural speech while controlling these features independently, even synthesising combinations of features not present in the training data. A model based on the one in Lyth and King (2024), called **ParlerTTS** (Lacombe et al., 2024), forms the baseline model used in the experiments in Chapter 13.

12.2.4 Evaluation

Evaluation of prompt-based TTS models often involves a mixture of objective metrics and subjective listening tests. Since many such models are trained on acoustic features mined from the training data, compliance with instructions can be directly evaluated by applying the same feature extraction pipeline to synthetic outputs (for example Yang et al., 2024; Lyth and King, 2024). InstructTTS additionally measures Mel-Cepstral Distortion (MCD), Voicing Decision Error (VDE) and F_0 Frame Error (FFE) for eval-

uating speech quality (Yang et al., 2024). Lyth and King (2024) employ three different neural networks to evaluate adherence to target gender, accent, and overall audio fidelity. Subjective listening tests are carried out to evaluate perceived naturalness (Yang et al., 2024; Lyth and King, 2024; Guo et al., 2023), style and emotion rendering (Yang et al., 2024; Guo et al., 2023) and overall prompt compliance (Lyth and King, 2024). Although there now exist some holistic benchmarks (Huang et al., 2025) which aim to evaluate prompt-based TTS models, these remain limited; different prompt-based models make different assumptions of what should be controlled, making cross-model comparisons complex.

12.2.5 Limitations and challenges

Prompt-based TTS models may offer an intuitive manner of control, but they are typically limited to utterance-level control only, where individual lexical units cannot be controlled in a targeted manner. Although prompt-based TTS models can generate output that adheres to the instructions, the user might wish to make targeted refinements to the generated speech. Some methods aim to provide a more targeted manner of control, but this is typically limited to a small subset of control features (e.g., Jin et al., 2024; Zhou et al., 2024),

The core challenge to prompt-based TTS is data availability. Few description-annotated speech corpora exist, so many approaches use LLM-generated annotations instead (e.g., Leng et al., 2024; Ji et al., 2024; Lyth and King, 2024; Seshadri et al., 2022). Although this supports model generalisation to free-form descriptions of the same features, control is still limited to the finite list of control features included in the description prompts used for training the model.

The next chapter presents a fine-tuning method for a prompt-based TTS model based on a SLM architecture. As such, the same inputs can yield diverse but coherent outputs. The proposed method exploits this model-generated output variation to expand the range of features available for control.

Chapter 13

A strategy for control feature discovery

The current chapter covers work from *RepeaTTS: towards feature discovery through repeated fine-tuning* (Sigurgeirsson and King, 2025), presented at SSW 13. Despite their user-friendliness, prompt-based models face limitations. On one hand, control is restricted to the features the model has learned during training. On the other hand, the same input can produce uncontrollable variation, influenced by unlabelled patterns present in the training data. In Sigurgeirsson and King (2025), we investigated a novel fine-tuning regime that addresses both issues simultaneously.

13.1 Research objective

Most Text-To-Speech (TTS) models offer no or limited control over how the output prosody is determined: the model produces a single, probable rendition given just the target text. While the models explored in this thesis provide some mechanisms for adjusting output prosody, they share an obvious limitation: the control they offer is limited to the features used for training the model. This constraint also applies to prompt-based models. Yet, these models can generate a range of plausible renditions for identical inputs, determined by corpus variation that the user has limited or no control over. To make the point clear, I give an illustrative example:

1. A prompt-based model has been trained on an expressive speech corpus with description prompts describing speaking rate and mean pitch.
2. A user prompts the model with a desired specification of those two features, e.g. *The speaker's pitch is high and they speak slowly*
3. Drawing on training utterances corresponding to the user's description, the model synthesises a probable rendition.
4. The user determines that the rendition is technically accurate with regard to the description prompt, but they wish the rendition were more expressive, exhibiting a more varied pitch range.
5. A typical expressive speech corpus will comprise speech that differs in pitch range. Yet, since the model is trained without labels describing this behaviour, the user cannot make this choice through the prompt.

Presumably, training utterances could be annotated with a comprehensive list of acoustic features derived directly from speech, enabling more comprehensive control over the model's output variation. However, a complete specification of a large set of acoustic features is unlikely to enable a user-friendly manner of control. Instead, we propose annotating the training utterances with features that capture variation across multiple acoustic dimensions, without requiring the explicit specification of each. To support this, we proposed exploiting the output variation of a prompt-based TTS model to find a small number of control features that account for a large proportion of the output variance:

> **RQ 13-1:** *Can control features be discovered from the output distribution of a prompt-based TTS model?*

To enable the controllability of a discovered feature, utterances from the initial training corpus are annotated with textual descriptions based on the analysis. We propose an iterative fine-tuning strategy to find a small number of control features that collectively account for a large part of the residual model output variance. After each fine-tuning round employing new descriptions, the output distribution of the newly fine-tuned model is analysed again, now exploring variation not accounted for by labels found in the previous round:

> **RQ 13-2:** *Can the model learn new control features through repeated fine-tuning?*

It should be noted that the nature of this work was exploratory. We did not know what sort of features would be discovered or how many iterations of the fine-tuning method would be required. The main goal was to find features that represented perceptually — as based on *our* perception — salient characteristics, which could be incorporated through fine-tuning to reduce overall output variance.

13.2 Chapter overview

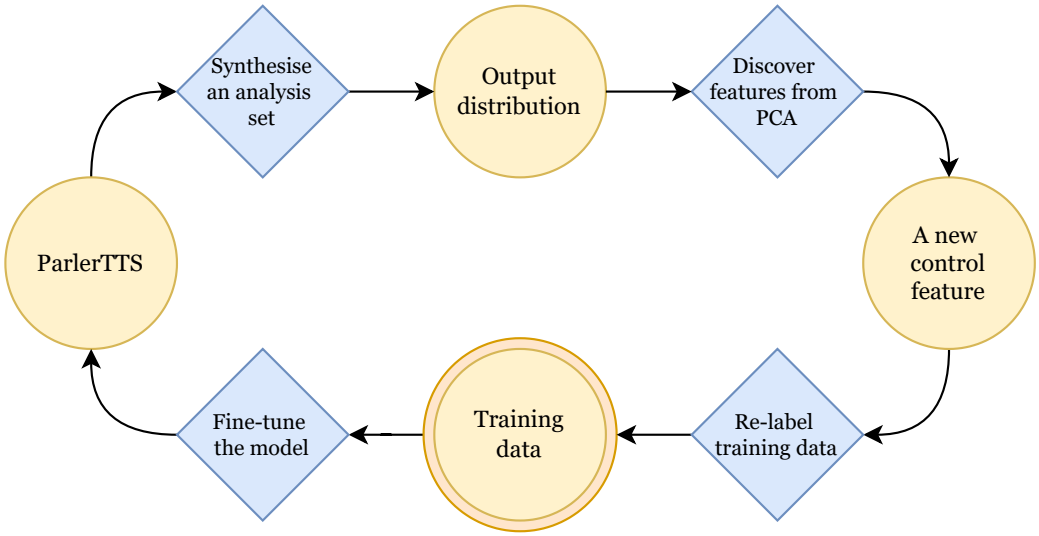


Figure 13.1: A flowchart illustrating the proposed four-phase cyclical fine-tuning approach, starting from the *training data* circle. Training data, labelled with natural language descriptions of control features, is used to fine-tune an initial *baseline* **ParlerTTS**. The fine-tuned model is then used to synthesise a large *analysis set* of utterances. A PCA of the analysis set aims to discover an additional control feature that reduces overall output variance. This feature is used to re-label the training corpus, and then the process is repeated.

The nature of the method is complex, involving multiple stages of fine-tuning and analysis. This section aims to give a clear overview of what follows. The four steps of the fine-tuning strategy are illustrated in Figure 13.1. The strategy begins with an initial *baseline* prompt-based model, a model trained on a corpus containing natural language descriptions of the *initial* control features. Then, the strategy aims to incorporate additional control features, determined through an analysis of the model output distribution. We based this analysis on Principal Component Analysis (PCA) of utterance-level representations (Wav2Vec2.0 (Baevski et al., 2020)) of synthesised samples. The rest of the current chapter is organised as follows:

Section 13.3: discusses the prompt-based TTS architecture, and checkpoints, used to create the *baseline* models for the fine-tuning experiments.

Section 13.4: explains how the PCA-based feature discovery is carried in the fine-tuning experiments.

Section 13.5: presents the results from the fine-tuning experiments for two different *baseline* **ParlerTTS** models.

13.3 Baseline models

13.3.1 Model architecture and data

We chose **ParlerTTS** (Lacombe et al., 2024), an implementation of the model proposed in Lyth and King (2024), as the model architecture for the current work. **ParlerTTS** is a Speech Language Model (SLM) architecture that predicts Descript Audio Codec (DAC) tokens [1], given the target text and the description prompt. During training, the target text tokens are prepended to the audio tokens. A frozen text encoder (T5 (Chung et al., 2024)) encodes the description prompt, which the model cross-attends to during audio code prediction. The description prompts are automatically generated to include information about the speaker's pitch (monotone or expressive), speaking rate and recording conditions.

The **ParlerTTS** authors have made several different model checkpoints available, including ones that employed a list of *known speakers* during training. This should enable the choice between known speakers during inference by including one of their names in the description prompt. However, we found that the synthesised speaker identity was frequently inconsistent with the mentioned speaker. Speaker consistency was a key requirement for our proposed method since we were interested in controlling paralinguistic features independently of the speaker's identity. Therefore, we constructed our own **baseline** models — by fine-tuning an existing **ParlerTTS** checkpoint — for the experiments and then confirmed successful speaker identity modelling.

We used an expressive training corpus for training the *baseline* models. Presumably, an explicitly expressive corpus features a broader range of prosody than an inexpressive one. We employed *Talromur-3* (Örnólfsson et al., 2024), an Icelandic, multi-speaker

[1] https://github.com/descriptinc/descript-audio-codec

emotive speech corpus. In *Talromur-3*, the same list of approximately 400 text prompts is spoken according to 5 different emotion labels: *happy*, *sad*, *angry*, *surprised*, and *helpful*[2]. The corpus's utterances are additionally labelled with an emotional intensity level from a 5-point Likert scale (*very low, low, medium, high, very high*). The same list of prompts was also spoken in a *neutral* emotion, without an intensity level.

The primary **ParlerTTS** checkpoints[3] that have been made available employ text tokenisation that does not support Icelandic text. Therefore, the *baseline* models were fine-tuned from a multilingual **ParlerTTS** model[4] instead. This multilingual model was originally trained on eight different European languages, none of which is Icelandic. Different from the original **ParlerTTS** model, this version was trained using LLaMA's tokeniser (Touvron et al., 2023) as a separate text prompt tokeniser. This tokeniser can be extended to other languages, such as Icelandic, using byte fallback (Kudo and Richardson, 2018). The multilingual **ParlerTTS** still employed a T5-based encoder (Chung et al., 2024) for encoding description prompts. So, the description prompts we used to train our checkpoints were written in English, as opposed to Icelandic.

13.3.2 The two baseline models

We created two *baseline* models, **T3-emotion**[5] and **T3**, by fine-tuning the multilingual **ParlerTTS** checkpoint. **T3-emotion** was fine-tuned with textual descriptions of the emotion category and emotional intensity, in addition to a description of the other features that the multilingual checkpoint was already trained on. The other model, **T3**, was fine-tuned using description prompts that included no information about the emotion category or intensity. The emotional labels are salient indicators of the prosody employed for a given sample in the training corpus. Therefore, for **T3** where these labels were not included, we expected that the proposed method would discover control features directly relevant to the unlabelled emotional categories.

We used the `Dataspeech`[6] repository to generate the description prompts for *Talromur-3*. `Dataspeech` employs `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023) to generate

[2] *Helpful* can be considered as *child-directed*.

[3] The `mini` and `large` series of checkpoints available at: https://huggingface.co/collections/parler-tts.

[4] https://huggingface.co/parler-tts/parler-tts-mini-multilingual

[5] T3 is in reference to the corpus employed, Talromur-3.

[6] https://github.com/huggingface/dataspeech

diverse natural language descriptions of training utterance features. We created a new instruction prompt to include information about the emotional content of the utterances for **T3-emotion**. A snippet of the prompt is shown below, while the full prompt is shown in Appendix F. The two baselines (**T3-emotion** and **T3**) were fine-tuned for 36 epochs on the Talromur-3 corpus.

Prompt Snippet for T3-emotion

```
You will be given 7 descriptive keywords related to an audio
sample of speaker_name's speech. These keywords include: The
gender (male, female), the level of reverberation ..., the
emotion of the speaker's voice. This could be one of Happy, Sad,
Angry, Surprised, Helpful or Neutral. This will also include
the intensity of the emotion, for example, high-intensity sad
emotion: the speaker is sad, and the intensity of the emotion is
high.

Your task is to create a text description using these keywords
that accurately describes the speech sample ... For example,
given the following keywords: 'female', ..., 'high intensity
angry emotion' a valid description would be: 'speaker_name
speaks very slowly but has a very animated delivery. She sounds
noticeably angry. The recording is noisy ...'
```

13.3.3 Baseline model evaluation

For the fine-tuning strategy to capture a large proportion of plausible model variation, the baseline output had to be acoustically diverse. We evaluated the diversity of the output by computing the mean pairwise cosine distance of Wav2Vec2.0 (Baevski et al., 2020) utterance-level embeddings of all speech samples in a given analysis set[7]. We employed a 300-million parameter multilingual version (Babu et al., 2022) of Wav2Vec2.0 to extract these embeddings. To make utterance-level comparisons, we took the mean of all predicted Wav2Vec2.0 embeddings for each synthesised utterance (the *summary embedding*). Lin et al. (2023) found that early layers of models like Wav2Vec2.0 correlate more with prosodic information structures than later layers. For Wav2Vec2.0 in particular, the first four layers were found to carry most of the prosodic information. Based on this, we chose Wav2Vec2.0's 4th layer to extract these embeddings. For a set of N summary embeddings, we used a *diversity score* to evaluate the overall diversity of an analysis set:

$$\text{diversity score} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} (1 - \cos(\mathbf{e}_i, \mathbf{e}_j)), \quad i \neq j$$

[7]The term *analysis set* refers to the set of model-generated utterances from which feature discovery is performed.

where $\mathbf{e}_i$ and $\mathbf{e}_j$ are the summary embeddings corresponding to the i^{th} and j^{th} utterance in the analysis set respectively. A *higher* diversity score corresponds to lower pairwise cosine similarity across the set.

The analysis set's diversity can be artificially increased by adjusting generation parameters — such as sampling temperature — to make the model more likely to choose less probable continuations[8]. But highly divergent sampling parameter values may result in reduced overall utterance quality. So, we contrasted diversity with *speaker similarity* and intelligibility in various settings to account for potential degradation in quality. Intelligibility was measured based on the Word Error Rate (WER) of transcriptions predicted by a Wav2Vec2.0-based Automatic Speech Recognition (ASR) system for Icelandic[9]. We used Resemblyzer (Wan et al., 2018) to extract speaker embeddings. 300 randomly sampled utterances of each ground-truth speaker were used to create the ground-truth mean speaker embeddings. Speaker similarity—the model's ability to consistently reproduce a target speaker identity—was evaluated based on the average cosine distance between speaker embeddings extracted from synthetic speech and speaker embedding means of ground-truth speakers.

We first determined suitable generation parameters (Section 13.3.3.1), then compared the model to ground-truth utterances in a broader context using the chosen parameters (Section 13.3.3.2). We performed this initial model analysis only for the **T3-emotion** baseline. We were interested in evaluating how the ParlerTTS output variation compares with the ground truth variation in general. **T3-emotion** was therefore more suitable than **T3** as variation can be explicitly induced by including emotion labels in the description prompts.

13.3.3.1 Choice of generation parameters

We evaluated relations between sampling parameter choices and diversity, intelligibility, and speaker similarity. We employed 100 utterances generated by **T3-emotion** for each corpus speaker using diverse description prompts and randomly sampled text prompts from an evaluation set.

Several generation-hyperparameters affect the range of model output variation: temperature (τ), k in top-k sampling, p in top-p sampling and the number of beams. During inference, we can control, using these parameters, how the model samples from the

[8] I.e, a continuation of speech tokens, given the ones already predicted.

[9] https://huggingface.co/language-and-voice-lab/wav2vec2-large-xlsr-53-icelandic-ep30-967h

distribution over possible continuations. We limited our parameter analysis to temperature and k, keeping p fixed at 0 and the number of beams at 1. Increasing temperature leads to less likely continuations being sampled more frequently, resulting in higher variation in the output. Increasing k forces the model to consider more candidates for continuations, so increasing k alongside an increase in temperature can further boost variation. While we aimed to generate diverse analysis sets, high values of these two parameters can also degrade the overall quality of the synthesised speech, as the model samples ever less likely continuations.

We measured diversity, intelligibility and speaker consistency for a $\tau \times k$ grid where values of τ are uniformly selected from $[1.0, 1.5]$ and values of k from $[50, 300]$. The results of this are illustrated in Figure 13.2. The third sub-figure shows what was expected: when these parameters increase, so does the diversity score of the analysis set. The figure also demonstrates the expected trade-off between diversity and quality of output. As τ and k increased, so did the measured WER of the system (first sub-figure). Similarly, the second sub-figure shows that speaker similarity drops substantially as either temperature or k increases. Variation in speaker identity will affect our analysis. Based on these results, we chose $\tau = 1.2$ and $k = 100$ for creating all our analysis sets; otherwise, we did not change the values of any other generation hyperparameters.

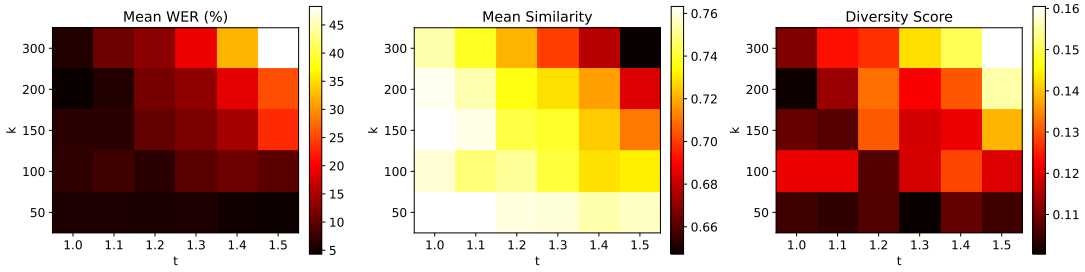


Figure 13.2: An overview of how the choice of temperature (τ) and k for k-sampling affects the output intelligibility, speaker similarity, and diversity.

13.3.3.2 Comparisons with ground truth utterances

Having determined suitable generation parameters, we compared 700 utterances generated by **T3-emotion** to 700 randomly sampled ground-truth utterances (100 per training corpus speaker). We compared per-speaker diversity, intelligibility, and speaker similarity between the ground-truth and the analysis set. Results are given in Ta-

ble 13.1. Our own initial observations of **T3-emotion** showed that the model is capable of producing highly natural-sounding speech while also adhering to the target emotion and speaker mentioned in the description prompt. However, we found that **T3-emotion** is significantly ($\alpha < .05$) less intelligible than ground-truth utterances according to a two-sample t-test ($t(700) = -15.2, p < .001$). We do observe that, as is common for models like **ParlerTTS**, the model occasionally hallucinates a repetition in the middle of a word. We hypothesised that this is the primary contributor to this decreased intelligibility. The synthetic data was also significantly more diverse than the real data ($t(700) = -18.9, p < .001$). There is a possibility that model-induced hallucinations may also have contributed to this effect. There was no statistical difference between the two sets in terms of speaker similarity ($t(700) = 0.0, p = 1.0$). Based on initial inspection, we determined that a speaker similarity of 0.75 or higher corresponded to perceptually high similarity to one of the real speakers. Several of the synthetic speakers were above this threshold, but *Ingrid* was chosen as the primary speaker in our experiments.

Table 13.1: Comparison of ground-truth (GT) and synthetic speech (**T3-emotion**) in an objective evaluation of intelligibility, speaker similarity, and diversity. Results are reported with standard deviations. It should be noted that WER scores are not normally distributed, so the standard deviation may not fully capture variability.

Spkr.	**WER (%)**		**Speaker similarity**		**Diversity score**	
	GT	synth	GT	synth	GT	synth
Astrid	2 ± 5	13 ± 17	0.78 ± 0.04	0.76 ± 0.05	0.13 ± 0.06	0.20 ± 0.06
Anders	4 ± 7	12 ± 15	0.71 ± 0.04	0.71 ± 0.05	0.15 ± 0.03	0.20 ± 0.08
Ingrid	4 ± 9	13 ± 19	0.77 ± 0.05	0.77 ± 0.05	0.14 ± 0.05	0.19 ± 0.06
Frida	5 ± 10	15 ± 18	0.72 ± 0.04	0.73 ± 0.05	0.14 ± 0.03	0.21 ± 0.11
Leif	3 ± 7	10 ± 13	0.75 ± 0.06	0.75 ± 0.04	0.12 ± 0.07	0.20 ± 0.08
Freya	5 ± 11	12 ± 15	0.75 ± 0.04	0.75 ± 0.04	0.14 ± 0.03	0.21 ± 0.09
Bjorn	6 ± 10	14 ± 16	0.71 ± 0.04	0.71 ± 0.05	0.12 ± 0.02	0.23 ± 0.10
All	4 ± 7	13 ± 14	0.74 ± 0.04	0.74 ± 0.05	0.13 ± 0.05	0.21 ± 0.10

We performed a secondary *neutral-speaking* trial, where **T3-emotion** was prompted with the *neutral* emotion label, to evaluate whether the speakers were consistently dissimilar from each other. We generated 1,000 utterances per speaker and determined, for each utterance, which ground-truth speaker it most resembled, based on cosine similarity. In all cases, the synthesised voices sounded most like the correct speaker. Figure 13.3 shows the similarity of synthetic *Ingrid* to all ground truth speakers in the corpus. As the figure illustrates, the similarity between the ground-truth and synthetic

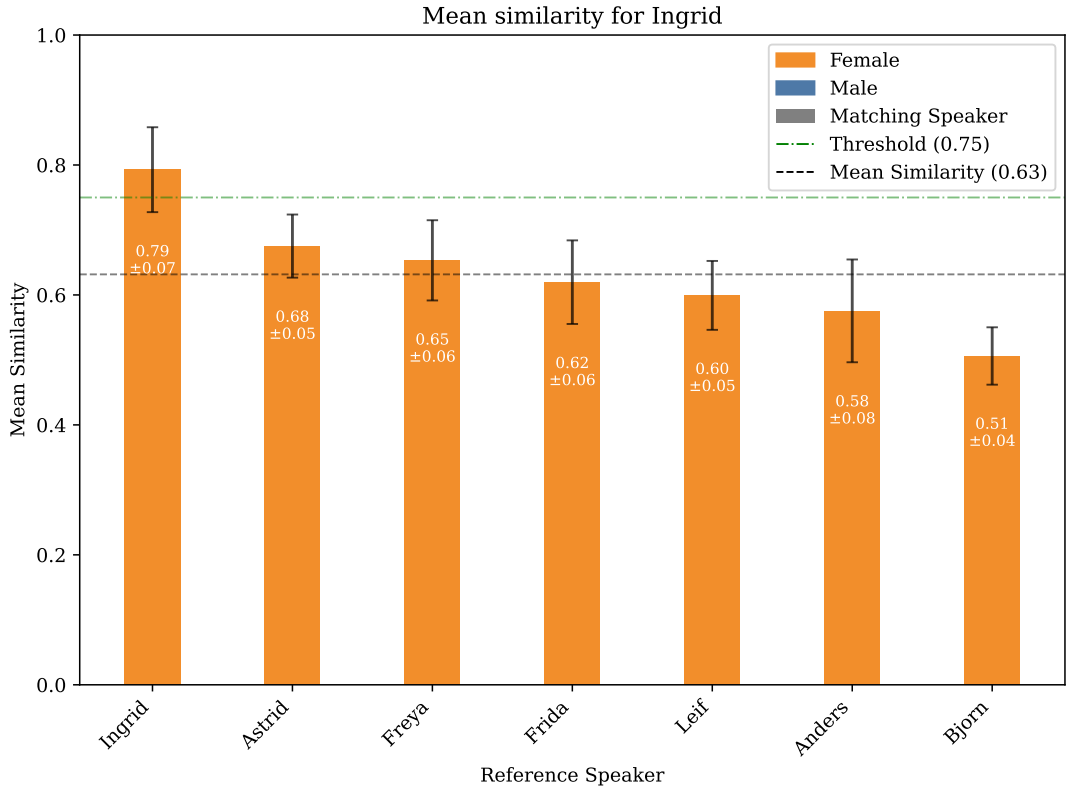


Figure 13.3: Mean speaker similarity of synthesised Ingrid to all ground truth speakers

Ingrid was the only one to cross the pre-determined 0.75 threshold.

We then investigated how well the model preserved a target speaker identity when prompted with different emotion labels. We measured the cosine similarity between ground-truth and synthetic speaker embeddings where the speaker and emotional labels matched. We found that **T3-emotion** actually preserved the speaker identity at a higher rate for emotional utterances when compared to neutral ones. The mean cosine similarity for all speakers was above the 0.75 threshold, as shown in Figure 13.4. This result suggests that the corpus speakers employed distinct prosodic characteristics for the different emotional categories, and that the model successfully learned this emotion-to-prosody mapping for each speaker.

Having found a good hyperparameter configuration for the baseline models and validated their ability to preserve speaker identities, we moved to the fine-tuning of the baselines. The proposed fine-tuning regime consists of two phases: (1) feature exploration, where new features are discovered from the output distribution of the *current* model, and (2) feature incorporation, where these new features are used to label the training corpus before fine-tuning the model again.

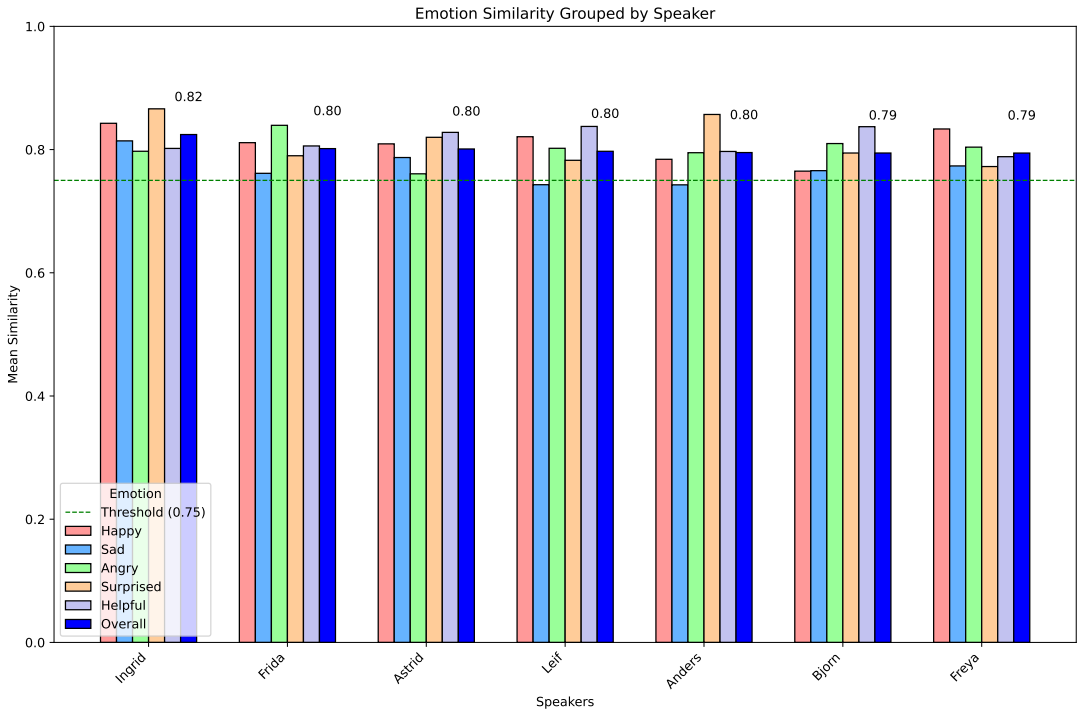


Figure 13.4: Mean speaker similarities between matching ground truth and synthetic speakers, broken down by emotional class.

13.4 Feature discovery and incorporation

We aimed to discover features that, through follow-up fine-tuning of the baseline **ParlerTTS** (**T3** or **T3-emotion**), could be controlled similarly to the other control features already learned by the baseline. The space of summary embeddings, which represents variation across a set of utterances, was employed to perform the feature discovery. At its core, the discovery method is based on a Principal Component Analysis (PCA) of summary embeddings across a large number of synthesised samples (the *analysis set*). After synthesising each analysis set, we projected the summary embeddings into three dimensions using PCA. We performed this projection to make feature exploration tractable: the original embedding space is too complex to support the manual feature analysis we employed to determine the characteristics of the feature yielded by the exploration.

We aim for each analysis set to be varied to support feature exploration. Given fixed sampling hyperparameters, **ParlerTTS** can still be explicitly made to generate variance in several ways: (1) through choice of target text, (2) choice of target speaker (if the model supports such a choice), and (3) other instructions included in the description prompt. Diversity in one or more of these inputs would naturally contribute to

the observed model variance for the analysis set. We wanted to determine which compositions of those variables would be most conducive for feature exploration. Several different compositions were initially evaluated for the fine-tuning experiments. We tested single-speaker and multi-speaker sets, as well as single- and multi-text sets, and sets with diverse description prompts. In total, over 100,000 utterances were generated across all analysis sets. A large portion of these utterances are later visualised in Figure 13.8 (p. 178). However, we found that the summary embeddings are inherently speaker- and text-dependent. These dependencies hold no matter which layer of Wav2Vec2.0 we used to generate the embeddings (see Figure 13.5). In our initial baseline evaluation (Section 13.3.3), the diversity scores were computed from an analysis set employing multiple speakers and texts, and therefore affected by these dependencies. But that evaluation was performed primarily to gauge the model's general abilities to replicate the corpus variance. For feature discovery, we were interested in features that did not reflect differences in speaker or text. Because of this, we focused only on *fixed inputs*: analysis sets were generated using the same target text, target speaker, and description prompt.

Since we limited our analysis to fixed inputs, we assumed that most of the observed variation would be explained by differences in uncontrolled variables, such as prosody. We expected variance stemming from suprasegmental categorical features (like speaking styles) to appear as distinct clusters in the PCA space. In contrast, continuous ones (like mean pitch) would manifest as gradients within those clusters. By manually inspecting the PCA projections of summary embeddings, we determined whether the observed variation resulted in either discrete or continuous control features, which were handled slightly differently. Given distinct clustering in PCA space, corresponding to a categorical control variable, we applied K-means (Lloyd, 1982) to compute mean Wav2Vec2.0 embeddings for each cluster. Each *training utterance* was then labelled given the label of the nearest cluster. In cases where the PCA revealed a continuous gradient instead, we discretised the synthesised samples into n bins along the principal axis of variation. As with discrete variables, we computed mean embeddings for these bins and assigned feature labels based on cosine distance. Similar analyses, based on PCA, have been carried out to analyse model variance (see, for example, section 3.2.3 in Karapiperis et al. (2024)). The flexibility of prompt-based control allows us to express these cluster or bin labels as natural language descriptions that reflect their observed acoustic characteristics.

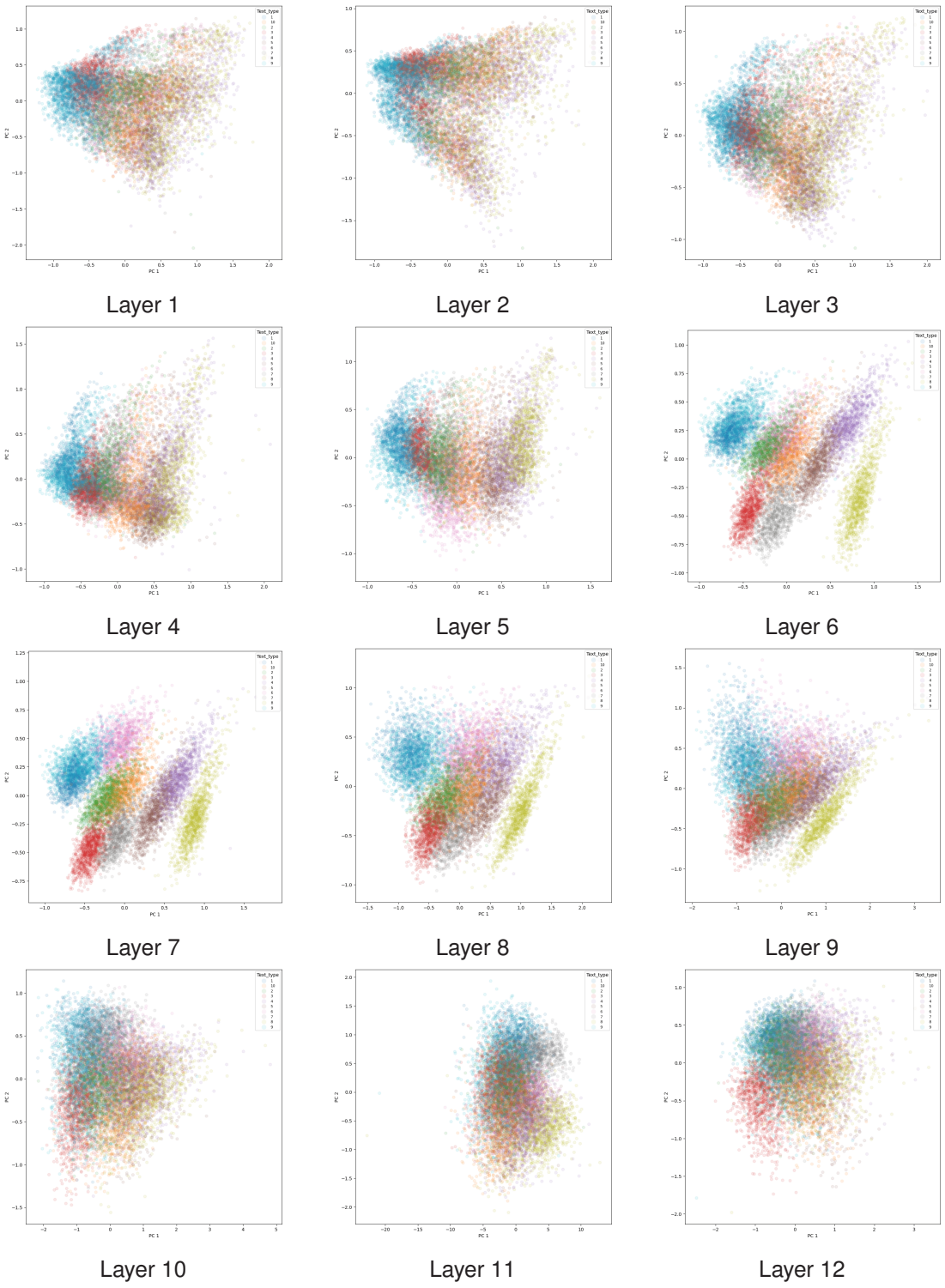


Figure 13.5: 2 component PCA visualisation of summary embeddings, generated by different layers of Wav2Vec2.0 (Baevski et al., 2020), for synthetic speech. Embeddings are coloured according to the 10 different target texts used to generate this particular analysis set.

The incorporation of a hypothetical feature

To illustrate how newly discovered features are incorporated into fine-tuning, take, for example, a hypothetical continuous feature **feature$_x$**. Based on PCA, three mean embeddings, corresponding to three locations on the axis of variation of this feature, have been created. It has been determined that the *magnitude* of this feature positively correlates with this axis. Hence, training utterances closest to the first embedding have the least amount of **feature$_x$** and those closest to the third have the highest. By matching training utterances based on distances to these embeddings, we can simply append *"The voice has low/medium/high intensity **feature$_x$**"*, to incorporate the features before fine-tuning the model again.

13.5 Fine-tuning results

The fine-tuning approach was first evaluated on **T3**, as a hypothetically easier test case: since **T3** was trained without emotional labels, generating an analysis set from fixed inputs could still reveal high prosodic variation corresponding to *hidden* emotional labels. In contrast, **T3-emotion** relied on explicit mentions of the emotional labels to generate emotional variation, making the feature discovery potentially more constrained. As previously stated, we expected the analysis of **T3** to reveal features directly related to the emotional labels, as these are salient cues of the prosody employed by the corpus speakers. **T3-emotion** was, on the other hand, used as a more challenging test case for the proposed method, in which we investigated which additional features the method would be capable of discovering.

In addition to Wav2Vec2.0 summary embeddings, we extracted a host of acoustic features for each utterance in each analysis set. We estimated the F_0 contours of synthesised speech using REAPER[10] and extracted functional features from the `GeMaps-v01b` feature set using OpenSmile (Eyben et al., 2010). We used these to look for correlations between the discovered control features and known acoustic features.

13.5.1 Fine-tuning of T3

We aimed to only introduce one new control label at a time before fine-tuning the model again. After each fine-tuning phase, we repeated the analysis to capture the remaining variation not yet explained by previously incorporated control features.

[10] https://github.com/google/REAPER

13.5.1.1 First fine-tuning stage

We created a single-text, single-speaker analysis set, using *Ingrid* as our target speaker, using the same, simple description prompt for all samples:

Initial description prompt

```
Ingrid sounds very clear and close to the microphone.
```

The analysis set contained 1,000 renditions synthesised using this prompt, allowing the model to sample whichever prosodic rendition it deemed probable for the inputs. Based on our observations, the analysis set featured a wide range of vocal renditions, reflecting both emotional class and intensity. PCA revealed that the first two principal components accounted for 42.7% of the explained model output variance, determined by a full-rank PCA. We estimated the overall variance present in the analysis set by calculating the normalized trace of its covariance matrix[11]. This quantity corresponds to the sum of variance across all embedding dimensions, providing a measure of the overall data dispersion. The estimated total variance of the set was 0.513. As a comparison, the estimated variance for an equally sized analysis set with no control over speaker, text, or description had an estimated output variance of 1.037.

We evenly sampled and listened to 30 generated utterances across the first principal component axis. We determined, from listening to the samples, that the first principal correlated with the emotional intensity of the rendition, from low to high intensity. There was no clear correlation with the emotional class itself. That is, utterances that we considered to be of high intensity included both samples that we perceived as *happy* and *angry*, for example. Therefore, we decided to simply label this discovered feature as *intensity*. We discretised the range into three bins and created appropriate labels for each (1: *low intensity*, 2: *medium intensity*, 3: *high intensity*). We created mean bin embeddings by taking the mean of 50 random samples from each bin, to re-label the training corpus according to the lowest cosine distance. Appropriate labels were then appended to the training corpus description prompts:

[11] Computed by summing the explained variance ratios of all components in a full-trace PCA of the set, normalised by dimensionality.

Description prompt with intensity label

```
Ingrid sounds clear and close to the microphone. Ingrid speaks
at low/medium/high intensity.
```

We can assess how well the new labels generalise across the training corpus by examining how the ground-truth emotional intensity labels — unseen by **T3** — are distributed among the three new categories. A close match between discovered intensity labels and ground-truth ones would suggest a high degree of generalisation. Table 13.2 lists which ground truth emotional *types* are paired with which cluster. We categorised any ground truth utterance with an emotional intensity higher than 3 as the *Emotion - high intensity* type and otherwise as the *Emotion - low intensity* type. Since neutral utterances in the corpus do not have an emotional intensity, they are kept separate. The table shows a substantial overlap of actual emotions across the three clusters, indicating only partial success in classifying emotional intensity based on the new labels.

Table 13.2: Classification of different ground-truth intensity types, based on labelling (*low*/*medium*/*high intensity*) from the first fine-tuning stage. *Emotion - high intensity* refers to ground-truth intensity labels above 3, and *Emotion - low intensity* to intensities below 4. A perfect classification would yield 100% on the diagonal and 0% elsewhere.

Ground-truth intensity	**cluster 1** (*low*)	**cluster 2** (*medium*)	**cluster 3** (*high*)
Neutral	59.3%	31.1%	9.6%
Emotion - low intensity	24.4%	48.3%	27.3%
Emotion - high intensity	5.3%	36.5%	58.2%

13.5.1.2 Second fine-tuning stage

We then fine-tuned **T3** again, now using instruction prompts modified to include the new control labels for intensity. We limited further fine-tuning to just *Ingrid* and carried out 40 epochs. The new prompt could now yield different emotional characteristics, based on the value of the new control label (*low/medium/high intensity*). We generated three analysis sets, one for each new control feature value, comprising 1,000 samples each. These three sets were analysed separately. We found that the first two principal components account for, on average, 23.7% of the total explained variance. The average estimated variance across the three sets was 0.413, indicating a reduction in model output variance as a result of the fine-tuning.

Interestingly, analyses of these sets repeatedly yielded two distinct clusters in PCA space, as illustrated in Figure 13.6 for one of the subsets. We performed a 2-component

K-means analysis on the synthesised embeddings. Through listening to samples, we observed that one cluster corresponds to the *Neutral* emotion label. At the same time, emotive speech was contained within the other, ranging from low-intensity to high-intensity as before. This separation was consistent across the three analysis sets.

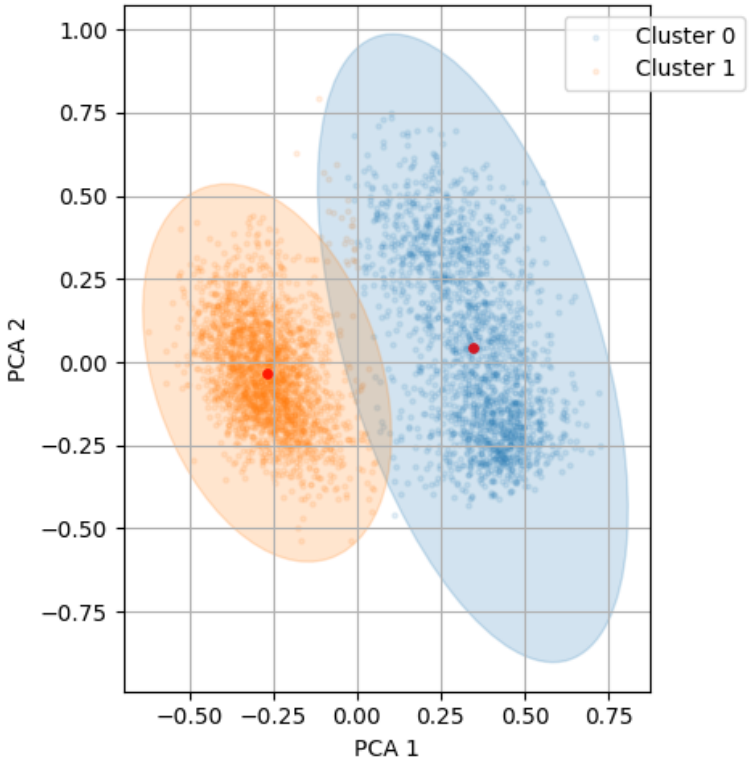


Figure 13.6: K-means clustering of projected embeddings from the *T3 - Ingrid* subset. The blue K-means cluster numbered 0[12] comprises emotional renditions, while cluster 1 contains neutral renditions.

Therefore, this second stage analysis found both a previously unseen feature (the *neutral* emotional class) and another one previously discovered (*intensity*). We created a mean embedding representing the neutral cluster (labelled 1 in Table 13.3) and two embeddings for high and low *intensity* utterances as before (labels 2 and 3 respectively). We then labelled the *Ingrid* subset again according to cosine distance to cluster means. We could have only labelled utterances assigned to the *neutral* mean and used the *intensity* labels from the previous stage for the remaining samples. But only a handful of utterances would now be assigned the previous *low intensity* label. So instead, we used the assignment produced by the current fine-tuning stage, overwriting the previous *intensity* labels.

[12] Confusingly, corresponding to *label cluster 1* in the labelling that follows

Description prompt with intensity or neutral label

```
Ingrid sounds clear and close to the microphone. Ingrid's voice
sounds neutral / a little intense / highly intense
```

Again, we evaluated which ground-truth emotional intensity labels were matched to each new label. The results of this are shown in Table 13.3. We saw an overall reduction in the confusion of neutral and non-neutral utterances, validating the clustering approach. We then initiated a third fine-tuning stage, now utilising the feature labels discovered during the second fine-tuning stage. Separation based on these new labels resulted in a further reduction in mean total variance, which was now down to 35.3%. Further fine-tuning stages would likely yield other control features. But, time constraints prevented further analysis of **T3**.

Table 13.3: Labels (*neutral*/*low intensity*/*high intensity*) assigned to ground-truth utterances after second fine-tuning stage, categorised by ground-truth intensity labels.

Ground-truth intensity	**1** (*neutral*)	**2** (*low intensity*)	**3** (*high intensity*)
Neutral	89.3%	5.2%	5.5%
Emotion - low intensity	12.3%	49.1%	38.6%
Emotion - high intensity	1.4%	33.3%	65.3%

13.5.2 Fine-tuning of T3-emotion

The results from Section 13.5.1 suggest that the fine-tuning approach can discover and incorporate some unlabelled emotional features in the training corpus. We started our analysis of **T3-emotion** with a single-speaker, single-text neutral-emotion analysis set with the aim of finding features not directly related to emotional content. We synthesised 1000 samples in this case, using the same target text (*Það var fallegt veður í gær og ég sá mörg dýr á leið minni í gegnum skóginn*[13]) and the same description prompt:

Initial description prompt for T3-emotion

```
Ingrid speaks in a neutral tone without any particular emotion.
Ingrid's voice is clear, and she is close to the microphone.
```

There are several plausible prosodic renditions for this text, but Figure 13.7 illustrates that the model is highly consistent in its intonation. The figure shows the mean and

[13]E: *The weather yesterday was beautiful and I saw many animals on my way through the forest.*

95% confidence intervals of the 1,000 different F_0 contours for the generated samples. As this overview illustrates, a large majority of the F_0 contours follow the same pattern.

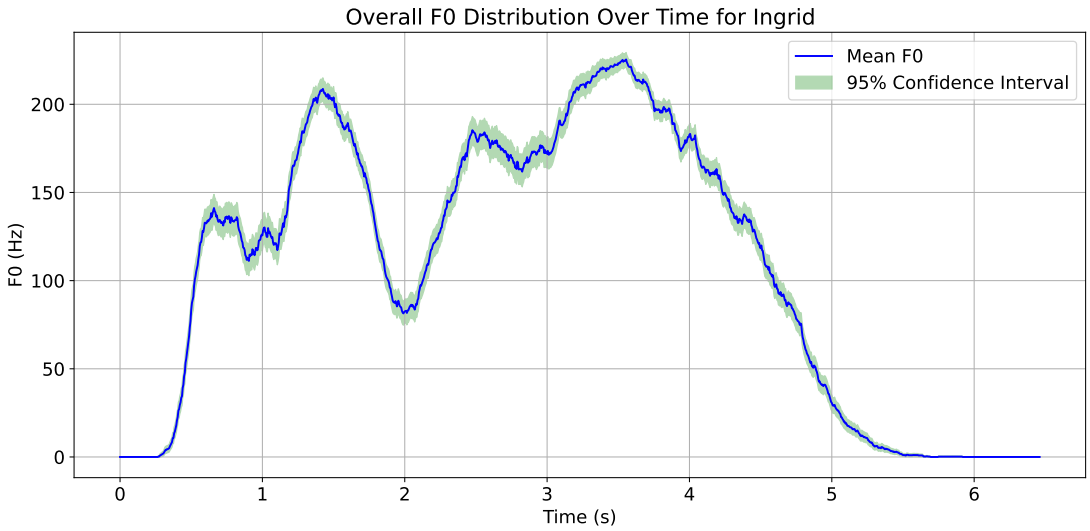


Figure 13.7: Overview of the 1000 predicted F_0 contours for the single-speaker neutral emotion, same-text analysis set. While the analysis set exhibits some variation, it is highly constrained as evidenced by the narrow 95% confidence interval (green shaded area).

We hypothesised that the model had learned a specific mapping for the neutral emotional label and would, therefore, not deviate too far from what it considers to be probable for that input description. When we compare this set to other renditions of the same text from more diverse analysis sets, we see that the neutral single-speaker set is very limited in variation, as illustrated in Figure 13.8. This figure is a PCA visualisation of all renditions of the same target text across different analysis sets generated during the course of this research. Projections of utterances from the initial neutral analysis set, *Ingrid - Neutral*, are highly consistent with each other compared to other sets. **T3-emotion** was trained on utterances labelled with description prompts that included information about emotional content. We assumed that omitting this information could yield greater output variation than if a specific emotional label were included in the description. So, we decided to prompt the model without any mention of emotion:

Updated description prompt for T3-emotion

Ingrid's voice is clear, and she is close to the microphone.

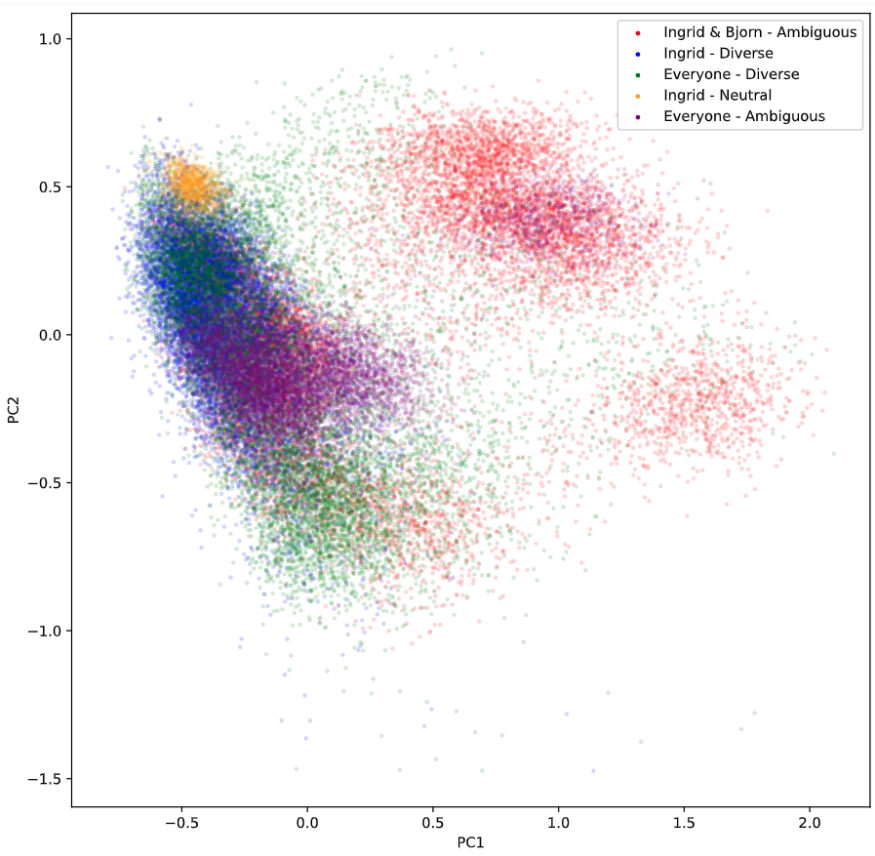


Figure 13.8: PCA projection of different **T3-emotion** renditions for the same target text, coloured by analysis set. The figure includes several analysis sets, differing in composition, which were synthesised during the course of this research, but we only employed analysis sets generated using fixed inputs.

Again, we used *Ingrid* as our target speaker. We observed a higher degree of output variation for this analysis set compared to the neutral one. We determined which of the `GeMaps-v01b` features best correlated with the first principal component. The results of this are shown in Figure 13.9. There is a strong negative correlation between features related to perceived loudness and the first principal component. By listening to samples evenly distributed across principal dimensions, we determined that the correlation with loudness is explained by variation in recording conditions across the Talromur-3 corpus.

The ParlerTTS checkpoint we employed to create our baselines already included features relevant to recording condition, namely *level of reverberation* and *noise*. The description prompt we used in the analysis controlled for both of these: "*Ingrid's voice is **clear**, and she is **close to the microphone***". Therefore, the variation resulting from differences in recording conditions cannot be attributed to an under-specification of features relevant to it. The initial attempt at feature incorporation did not yield

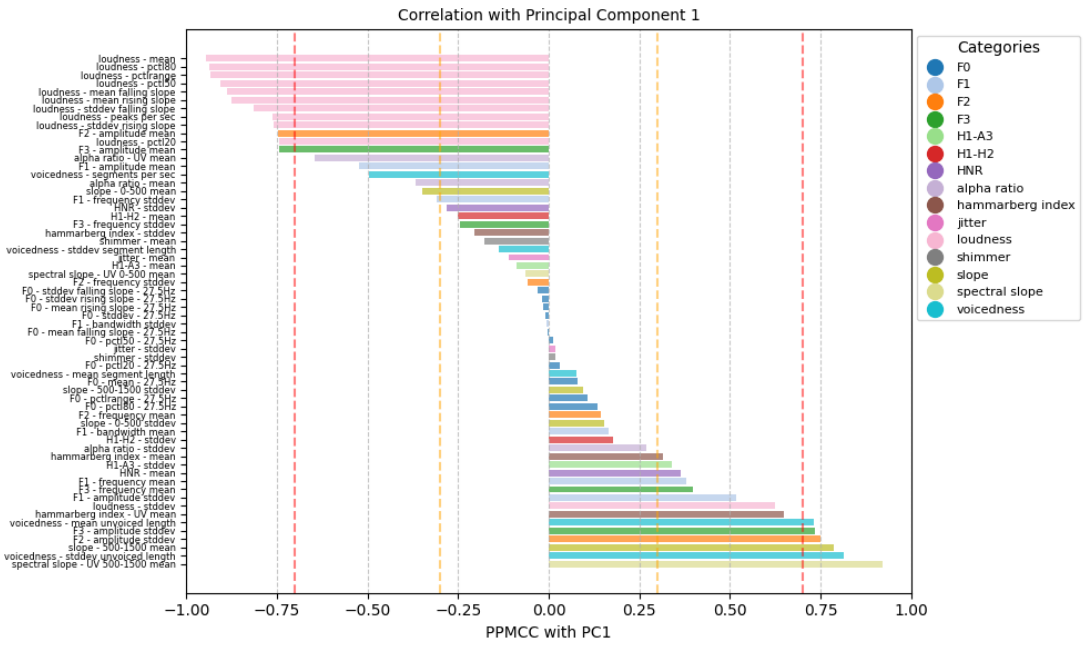


Figure 13.9: Correlation of `GeMaps-v01b` features with principal component of **diverse-ingrid**

a feature directly related to prosodic differences. However, the identified feature — *recording condition* — is still valuable, as the features intended to control for it show substantial variation. However, due to time constraints, we could not proceed to the next fine-tuning stage for **T3-emotion**.

13.6 Conclusion

We asked **RQ 13-1**: ***Can control features be discovered from the output distribution of a prompt-based TTS model?*** And then **RQ 13-2**: ***Can the model learn new control features through repeated fine-tuning?***. In addressing these questions, our method, as applied to **T3**, demonstrates a significant and novel capability: the discovery of previously unlabelled prosodic features. By systematically analysing thousands of samples generated from identical inputs, we were able to identify and cluster distinct prosodic variations. We then successfully integrated these latent features into the model using a secondary fine-tuning process, leveraging simple PCA-based clustering of utterance-level embeddings. This constitutes a powerful proof-of-concept for an unsupervised approach to control-feature discovery.

While his method proved effective with **T3**, a different outcome was observed with the other baseline, **T3-emotion**. Here, the method's inherent sensitivity to all sig-

nal variation, a key component of its power, instead captured artefacts related to the recording environment rather than prosodic features. This result, far from being a failure, is a crucial finding that highlights the method's capacity to identify even subtle, non-verbal variations. With additional fine-tuning — a step unfortunately limited by time constraints — we are confident this environmental noise could be isolated and controlled. This would enable subsequent stages to focus exclusively on discovering more meaningful prosodic features, further validating the methods' potential. To mitigate this in future work, alternative utterance representations, such as those that are less sensitive to non-verbal cues than the mean embeddings from multilingual Wav2Vec2.0, could be explored to refine the approach.

A core advantage of our proposed fine-tuning regime is its ability to generate and contrast thousands of prosodic renditions for fixed inputs. This overcomes a limitation of current research: the lack of large-scale, ground-truth datasets with perfectly matched speaker and text conditions. By fixing these variables, our analysis isolates and focuses solely on prosodic differences. While this experimental design confined our downstream fine-tuning to a single speaker, limiting our ability to test generalisation across speakers, the results of the feature labelling across different texts in the **T3** training corpus are promising. As shown in Tables 13.2-13.3, the discovered labels demonstrate generalisability across diverse texts. For instance, our method correctly classified 89.3% of neutral ground-truth utterances, despite emotional labels not being used in the initial training of the **T3** baseline. This high classification rate, while not the primary goal, serves as an indicator of the method's potential to uncover and meaningfully categorise latent prosodic features.

Chapter 14

Discussion

Compared to Acoustic Feature Control (AFC) control and reference-based control, prompt-based control can be considered more user-friendly; it's presumably easier for most people to describe the desired vocal behaviour using free-form text descriptions than with a reference utterance or a highly-specific description of acoustic features. But, like other control methods for Text-To-Speech (TTS), prompt-based control is limited to features that have been included in the training of the TTS model. This limitation is perhaps especially restricting for prompt-based models compared to the other control methods covered in this thesis. Prosody Transfer (PT) models cannot reliably produce prosody beyond the variance observed in the training corpus. However, using a reference within the training corpus range, the PT reference-based control enables the description of how prosody changes over time — something that prompt-based models typically do not provide. Similarly, an AFC model offers semi-continuous control of different acoustic features, whereas prompt-based models rely on discrete feature bins. As a result, prompt-based models face limitations not only in the breadth and statistical range of the control features they are trained on, but also in how users can specify their desired values. Consequently, there is a trade-off between user-friendliness and the specificity of control options.

That is not to say that one of these qualities of control is more important than the other. Different applications may require varying levels of specificity, and the required user-friendliness depends on the expected end-user. However, in the current section, we investigated whether this trade-off can be better balanced for prompt-based models. The work presented in Chapter 13 aimed to address this by exploiting the uncontrolled

variation exhibited in the model output for fixed inputs. The results demonstrate that the proposed method can discover salient prosodic features within the training corpus. However, they also reveal that the method is sensitive to any variance in the training data, which may complicate the discovery of prosodic features. The incorporation of discovered prosodic features, through fine-tuning, results in a reduction in overall model variance, indicating that the model has learned to effectively control them.

Chapter 15

Conclusions

Any text can correspond to multiple different spoken renditions. Often, knowledge of the conversational context is required to determine which one is most suitable. This variability poses a problem for Text-To-Speech (TTS) models, which must make this determination given limited contextual information. As a result, many TTS models produce a single most probable rendition, which may not be appropriate when expressive speech is anticipated. To mitigate this *one-to-many* problem, many TTS models make speech generation dependent on an additional conditioning signal. Variation in this signal allows for generating different renditions of the same text. Controlling the conditioning signal is what makes TTS models controllable. This form of control is very flexible, and different modalities can be used to condition the model. This thesis has discussed three methods, different in modality, for controlling expressive TTS.

As previously discussed in Chapter 1, the suitability of different control strategies comes down to the target task and the limitations specific to each method. Therefore, instead of suggesting a single best control strategy, this thesis aimed to research how the limitations of each method can be addressed (**RQ 1-1**) and in which circumstances each method is the most suitable (**RQ 1-2**). The following section provides a summary of the methods researched in the thesis to answer these questions.

15.1 Method summary

In Part I we investigated reference-based TTS models, focusing on Prosody Transfer (PT) models. They can generate multiple, prosodically-different renditions of the same

target text by conditioning generation on a supplied reference utterance. These models learn a latent reference space during training, which can be sampled from during inference. But finding an appropriate reference utterance is non-trivial. A reference utterance from a different speaker leads to *source-speaker leakage* (Sigurgeirsson and King, 2023). *Feature entanglement* (e.g., Skerry-Ryan et al., 2018; Battenberg et al., 2019) is a concern for different-text PT. Both greatly complicate the choice of an appropriate reference. Feature entanglement is a persistent issue in other reference-based applications as well. When applied to voice-cloning, the model fails to disentangle speaking style and speaker identity (Chapter 6).

In Part II we turned to Acoustic Feature Control (AFC) models, which learn to separately predict controllable acoustic features without explicit conditioning. Compared to reference-based TTS models, AFC models offer an interpretable form of control, since they allow for the temporally-precise adjustment of different acoustic features. The AFC model architecture enables precise control because it allows the user to control different segments of the output separately. In contrast, reference-based models typically influence the generation of the entire utterance. Limiting control to a small, but salient, set of acoustic correlates of prosody (F_0, energy, and duration) allows the model to generate prosodically-different and more context-appropriate renditions than a model employing unadjusted features (Chapter 9).

Lastly, Part III covered prompt-based models. These models, which learn a *prompt-to-acoustics* mapping from a prompt-labelled corpus, are controlled using a natural language description of the desired output. Therefore, users may find them very user-friendly, as they require neither a reference utterance for conditioning nor detailed knowledge of how specific acoustic features relate to the desired vocal rendition. But, there is a trade-off between user-friendliness and the specificity of control. Typical prompt-based models do not allow for targeted control of specific segments in the utterance, for example.

The suitability of a control method is highly dependent on the specific task the TTS system performs. So, the objective of this thesis was not to answer the question "*What is the best control method for TTS?*". Instead, the research presented in this thesis has examined these three different means of control and proposed modifications to each one to address their characteristic flaws. Reference-based models struggle with disentangling features. So in Part I, I proposed a novel training regime for a PT model to mitigate feature entanglement (Chapter 5). **FastSpeech 2**, an example of the AFC

models covered in Part II, technically allows for arbitrary control of specific acoustic features. But the modelling resolution is not conducive to control. Therefore, I proposed a more suitable control mechanism (Chapter 9), which was evaluated in two separate studies. Prompt-based models are limited to the control features on which they are trained. This limitation is, of course, true in general, but I argue that this is particularly limiting for prompt-based models (Chapter 14). To explore additional control features, a fine-tuning approach was described in Chapter 13.

15.2 Method comparison

Although this thesis does not explicitly compare the effectiveness of each control method, my research allows for a comparison of certain aspects of their controllability. The three control methods vary in several ways, but here I compare them in terms of their *interpretability*, *responsiveness* to the conditioning signal, *specificity*, and *usability*:

> **Interpretability**: To what degree can a user understand the relationship between the conditioning signal and the corresponding output?
>
> **Responsiveness**: Can features be controlled gradually? Do many different control inputs yield the same effect? Does a single control input yield many different effects?
>
> **Specificity**: What control resolution does the method offer? Can different segments in the utterance be controlled differently?
>
> **Usability** (i.e., "*ease of use*"): How easy is it for a non-expert to interact with the control strategy? Can the model be used without a conditioning signal?

Each method can be described by a *controllability profile*, indicating its performance in these four characteristics. Drawing on the differences and limitations of the three control methods, I now provide recommendations for the appropriate use of each strategy.

15.2.1 Reference-based models

PT models aim to transfer the reference prosody to the target text in a holistic manner. So, using a suitable reference utterance allows for controlling how the prosody devel-

ops throughout the output utterance. Certain aspects, such as speaking rate, can be gradually modulated; however, this requires finding or recording references that correspond to this gradual difference. Providing a reference utterance for target prosody may be more intuitive for certain applications than requiring a detailed specification of acoustic parameters for each phone, for example. But, reference-based models do not support isolated control of parts of the output utterance, like words or phrases, because the whole utterance is affected by the reference. Similarly, PT models do not allow for independent control of specific acoustic correlates of prosody. The reason for this limitation is that reference-based control learns to model all signal variation not already predicted by the input text, in a single, unified representation. So even if reference-based models can control several acoustic correlates of prosody jointly, they do not support isolated control of each one. Although they provide effective means for controlling speech variation, the reference embeddings lack interpretability. Results presented in Chapter 10 indicate that embedding similarity is a poor indicator of perceived prosodic similarity, suggesting that the latent reference space is not well-structured — a key indicator of the interpretability of learned representations.

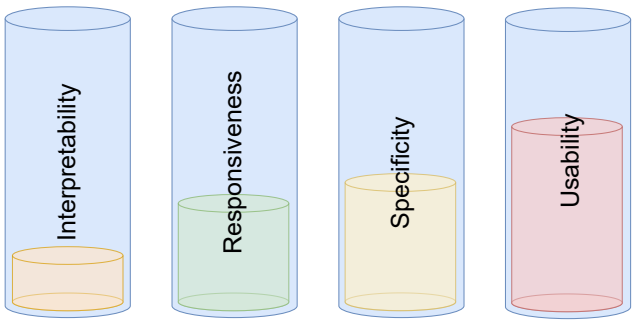


Figure 15.1: The controllability profile of reference-based models

PT models do not learn transferable representations of prosody. The nature of these models requires them to be trained under the same-text, same-speaker condition, resulting in feature entanglement. As Chapter 6 demonstrates, feature entanglement is a prevalent problem for other types of reference-conditioning as well. Feature entanglement further complicates the choice of acceptable references, since many references are invalid for conditioning in each case. Therefore, I advise against using reference-conditioning for different-text and different-speaker prosody transfer. Reference-conditioning can still be deployed successfully for alternative tasks, as evidenced by prior work covered in Chapter 3, and for constrained PT (e.g., voice puppetry).

15.2.2 Acoustic feature control

There are several key benefits to AFC. First, this approach is highly interpretable because the user can make modifications in terms of known and interpretable acoustic features. This interpretability also makes them excellent tools for analysis, as demonstrated in Part II, where AFC models were employed to study other forms of TTS control. Second, AFC-models enable the targeted control of acoustic features, allowing the user to apply them separately to target specific lexical units. This targeted capability enables more nuanced control than the other two control methods, which aim to control the entire utterance through a single conditioning signal. Also, in contrast to the other two methods, AFC models do not require a conditioning signal to synthesise; an AFC model predicts initial settings of control features, which are employed when control inputs are omitted. Therefore, AFC control inputs are optional and can be provided on a case-by-case basis. Lastly, AFC-models offer a highly flexible manner of control, which can be employed for a diverse range of tasks without requiring task-specific, labelled data to support the task.

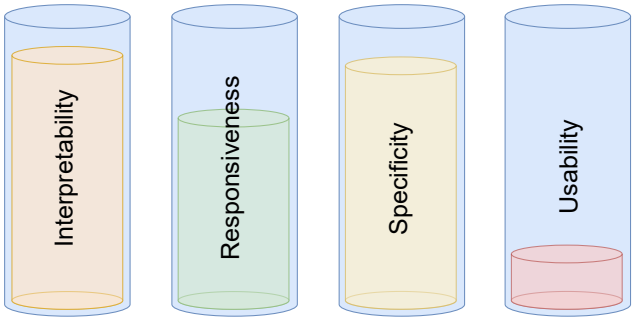


Figure 15.2: The controllability profile of AFC models

But results in Chapter 10 showed that it took Human-in-the-Loop (HitL) participants, on average, approximately two minutes to complete each control trial. So, despite the simplifications employed to make the task easy, substantial effort is required to control TTS models in this manner. Based on this summary, AFC controls are good candidates for tasks where nuanced changes to the output speech are anticipated, and less ideal for tasks where a more user-friendly strategy is required. Since AFC models do not always require control inputs to produce speech, they are a good candidate for use cases where only occasional modifications are needed.

15.2.3 Prompt-based control

In comparison to AFC and reference-based models, prompt-based TTS models provide a more user-friendly approach to TTS control. Describing the speech using natural language does not require expert knowledge, an appropriate reference for conditioning, or the careful control of acoustic features. Therefore, prompt-based models may be well-suited for applications where non-experts are expected to interact with the model. Prompt-based models typically do not offer targeted control, whereas both AFC models and reference-based models can make different adjustments to different parts of the output. Prompt-based models provide limited support for controlling the *level* of acoustic features, whereas the other two methods offer a more flexible approach. However, compared to PT models, prompt-based models enable the user to describe differences in just one control feature, without specifying any others.

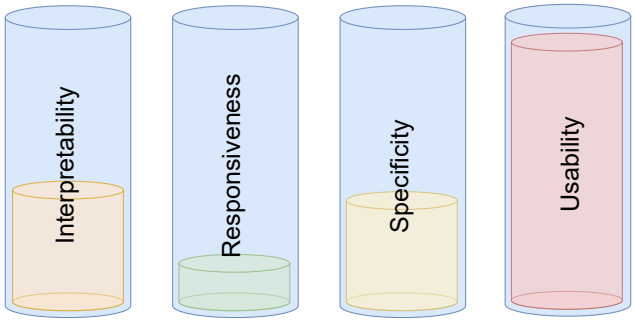


Figure 15.3: The controllability profile of prompt-based models

Prompt-based models based on Speech Language Model (SLM) architectures are non-deterministic: the same conditioning input can yield different effects in the output. Similarly, different description prompts may yield very similar effects. Like reference-based models, conditioning is performed using a learned representation that affects the whole output utterance. From this perspective, prompt-based models lack interpretability.

Therefore, prompt-based models are not ideal candidates for applications that require nuanced and targeted modifications. Because of their usability, they are better suited than the other two control methods for applications where non-experts are likely to interact with the model.

15.3 Final remarks

Much of the research presented in this thesis focused on enabling TTS models to generate expressive speech, in contrast to their typical tendency to produce inexpressive, *average* prosody. The period during which this thesis was developed coincided with rapid development of TTS. There are now many different TTS models that can generate remarkably realistic, high-fidelity speech. These models are easier to deploy and extend to different tasks than their predecessors. Not only do they generate highly realistic speech, but they can also produce varied prosodic renditions, express a wide range of emotions, incorporate non-verbal content, and generate very realistic dialogues; all of which contribute to their perceived naturalness. Given that today's state-of-the-art technology can already generate realistic and expressive speech, one might wonder whether the different control methods presented in the thesis remain relevant today.

The answer to that question is **resoundingly yes**. Despite rapid progress, TTS models still lack access to the context required to determine a suitable prosodic rendition in many cases. Control methods employing conditioning signals are still in use and are likely to continue for the foreseeable future. However, let's consider a distant future scenario where TTS models have become *black boxes* that wouldn't require a conditioning signal. In most cases, they could infer an anticipated prosodic rendition from limited contextual information. We wouldn't know how this complex system works; we would just know that it works in most cases. Because errors would occur so infrequently, we would have a limited understanding of why and how the system produces them. Then, more than ever, would it be crucial to regain control of the system to correct an unexpected output.

This is, of course, true for many other fields of research that focus on generative Machine Learning (ML) approaches. As generative models become more powerful, they require less explicit instruction for generation. Because means of control are required less frequently, they are phased out. As this evolution progresses, the model generally improves, but the previously necessary means of control are no longer available. While the risks implied by the loss of control of a TTS model may not be existential, there is nonetheless a risk to be considered. Speech is, at the end of the day, a social activity that evolves with us and is shaped by how we use it. However, today, more and more interactions are taking place between humans and machines. Now that TTS models have evolved from performing limited functional tasks to playing an active so-

cial role in spoken communication, some have begun to wonder whether the machine will eventually shape how humans speak (Székely et al., 2025). This is a genuine risk that should not be underestimated. Since the beginning of speech synthesis research, the relationship has been the other way around: the machine is expected to imitate humans. Von Kempelen's apparatus, the first major step toward speech synthesis, was founded on this very principle (Kratzenstein, 1781). Effective control methods are, therefore, not only practical tools that assist TTS systems in predicting speech. They represent an effort to maintain an understanding of how complex TTS models work and to preserve human influence over the speech generation process.

Appendix A

Pilot study: a parallel prosody transfer model

Preprocessing	Sample rate: 22050, Hop: 256, Win: 1024 Mel bins: 80, f_{min}: 0 Hz, f_{max}: 8000 Hz
FS2 Encoder	Dim: 256, Hidden: 128, Blocks: 4, Heads: 2, Dropout: 0.2
FS2 Decoder	Hidden: 256, Blocks: 4, Heads: 2, Dropout: 0.2, Postnet dim: 80
Variance Adaptor	Layers: 2, Filter: 256, Kernel: 3, Dropout: 0.5, Act: Rectified Linear Unit (ReLU)
Reference Encoder	Conv. layers: 6, GRU dim: 128
Speaker Encoder	Hidden layers: 1, Dim: 64
Optimizer	Adam (β_1=0.9, β_2=0.98, ε=10^{-9}) Batch: 16, Warmup: 4000, Grad clip: 1.0

Table A.1: Model hyperparameters for the proposed model

Appendix B

A training strategy for feature disentanglement

B.1 Participant instructions for first listening test

In this survey, you will rate the similarity of the prosody of a reference speech sample compared to different candidate speech samples. You will first listen to the reference. Then you will play one sample at a time and rate the similarity of the prosody of the sample, compared to the reference. Please note that what is spoken in the reference might be different from what is spoken in the samples compared.

When rating the similarity of prosody, please consider the following: 1. How pitch changes, i.e. rises and falls, throughout the sample 2. Stress put on each word or syllable (in terms of loudness or pitch change) 3. Speaking rate and how it changes throughout the sample 4. The length of pauses between words

In this survey, most audio samples have been edited in such a way that it becomes hard to tell what is being said. We say that these samples are muffled.

B.2 Participant instructions for second listening test

B.2.1 Prosodic similarity

In this section, you will rate the similarity of the prosody of a reference and 5 different samples.

You will first listen to a reference. Then you will play one sample at a time and rate the similarity of the prosody of the sample, compared to the reference. Please note that what is spoken in the reference might be different from what is spoken in the samples compared.

You will rate each sample on a 100-point scale from 0, being least prosodically similar, to 100, being prosodically identical.

When rating the similarity of prosody, please consider the following:

1. *How pitch changes, i.e. rises and falls, throughout the sample*
2. *Stress put on each word or syllable (in terms of loudness or pitch change)*
3. *Speaking rate and how it changes throughout the sample*
4. *The length of pauses between words*

Please ignore pronunciation issues, speaker identity and audio quality differences between the samples.

B.2.2 Speaker similarity

In this section, you will rate which of the two speakers sounds most similar to a reference speaker.

You will first listen to a reference. You will then listen to two voice samples. Then you pick the sample of the speaker you think most resembles the speaker of the reference. Please note that what is said in the reference will not match what is said in the samples.

Please ignore pronunciation issues and audio quality differences between the samples.

B.2.3 Naturalness

In this section, you will rate the naturalness of speech.

You will first listen to a speech sample. Then you will rate the naturalness of speech on a scale from 1 (completely unnatural) to 5 (perfectly natural).

Appendix C

The role of style in speaker identity judgements

C.1 Instructions for listening tests

C.1.1 First listening test

You will be played a series of voice recordings and asked to rate each along specific scales related to your perceptions about the speaker's voice. We are interested in your intuition (gut feeling) - so don't overthink.

For some voice clips, you will be asked to rate whether the voice of the speaker 'sounds gay' on a 7-point scale (1 - definitely sounds straight; 4 - neither straight nor gay; 7 - definitely sounds gay).

For other clips, you will be asked to rate how natural you perceive the speaker's voice to be on a 5-point scale (1 - Bad, 5 - Excellent).

The clips are of varying duration, please listen to the entire clip.

C.1.2 Second listening test

You will be played a series of voice recordings and asked to rate each along specific scales related to your perceptions about the speaker's voice. We are interested in your intuition (gut feeling) - so don't overthink.

For some voice clips, you will be asked to rate whether the voice of the speaker 'sounds

gay' on a 7-point scale (1 - definitely sounds straight; 4 - neither straight nor gay; 7 - definitely sounds gay).

For some clips, you will be asked to rate how natural you perceive the speaker's voice to be on a 5-point scale (1 - Bad, 5 - Excellent).

For others, you will be asked to rate how similar you perceive the voices of speakers to be on a scale from 0 to 100 (0 - completely different, 100 - exactly the same)

The clips are of varying duration, please listen to the entire clip.

Appendix D

A human-in-the-loop approach to improving cross-text prosody transfer

D.1 Target texts

Target texts used to create different-text Prosody Transfer (PT) stimuli:

- *The time passed very slowly.*
- *She immediately liked him.*
- *Gabriel, will you stay?*
- *Such a sad loss!*

D.2 Reference texts

Texts spoken in references used to create different-text PT stimuli:

- *And walking a few yards forward, while they talked together, soon made her quick eye sufficiently acquainted with Mr. Robert Martin.*
- *I have heard him express himself so warmly on those points!*
- *God's sake, yes–I am come to that low, lowest stage–to ask a woman for pity!*
- *Then where are you supposed to be getting the child?*
- *It is such a pretty charade, my dear, that I can easily guess what fairy brought it.*

Appendix E

Controllable speaking styles using a large language model

E.1 Instruction prompt

Listing: Instruction prompt and few-shot examples used in Sigurgeirsson and King (2024)

```
Could you help me with generating speech? I am trying to select the most
appropriate values of the following prosodic properties: duration,
energy and pitch.

Given some target text and a target speaking style. Assume that all the
speech is a part of a dialogue. Assume that I already have the speech
for the target text, but I can change the prosodic properties in a
relative manner. I will give you:
1. Either a previous line in the dialogue or a target speaking style
2. The target text

And I want you to tell me:
1. How to change the pitch, energy and duration, in general, for the
target text. We can increase or decrease these values as long as
they don't go out of the range that is considered normal for the current
speaker. So, for each of those attributes, tell me how much to change
them: (0: the standard value, -5: the minimum value,
5: the maximum value). A positive value for duration means a slower
speaking rate, and a negative value means a faster speaking rate. You
should select these values appropriately, given either the target style
or the dialogue context. These values should not be used to emphasise
specific words in the sentence.

2. How prominent is each word on a scale from 0 to 5 (0: standard
prominence, 5: maximum prominence). Note that this value should
reflect on how salient the word is in the given sentence, so most
words should probably have a value of 0. Salient words should
then have a value ranging from 1-5, considering how salient they
are in the sentence.

Instructions:
1. Report the sentence-level change of pitch, energy and duration
in a separate table. Pitch, energy and duration should be columns
```

```
in the table. Pitch first, then energy and finally the duration
2. Report the prominence level of each word in a separate table.
Remove any punctuation as we do not have to predict prominence for
those symbols.

Here are templates for the two tables:

The sentence-level attribute table:
|Pitch|Energy|Duration|
|---|---|---|
|the chosen pitch value|the chosen energy value|
the chosen duration value|

The prominence table (example sentence: he had a big house):
|he|had|a|big|house| ...
|---|---|---|...
|the chosen prominence level for word 1| the chosen prominence
level for word 2| ...

Rules you must follow:
1. All words in the target text should be included in the
prominence table.
2. Do not write any text above or between the tables. You should
include your reasoning after the tables.
3. You must use the table templates.
4. Your choice of parameters and reasoning has to make sense. For
example, you couldn't suggest a decrease, but then motivate that
choice by saying an increase is appropriate
6. Make your suggestions independently of who it is that is speaking.
The only thing that matters is that they are a native US speaker.

Your first example is the following:
# example 1
Target speaking style: Narration
Target text: "The old woman walked through the door."

These are the changes I would suggest:

Pitch   Energy  Duration
-1      -1      2

Word    Prominence
The     0
old     0
woman   0
walked  0
through 0
the     0
door    1

Explanation:
We make the overall duration slightly longer since that is fitting for
narration. We lower the pitch and energy for a calm-sounding voice,
which is fitting for this neutral target text. We make the word "door"
prominent since it is the most important word in the sentence.

# example 2
Target speaking style: Inspirational speech
Target text: "You can do it!"

Here is how I would change the prosody:

Pitch   Energy  Duration
2       3       -2

Word    Prominence
You     1
can     3
do      0
```

```
it      1

Explanation:
We make the overall duration shorter since the speaker is probably
excited,given the target speaking style. We further raise the pitch and
energy for a more excited-sounding voice, which is fitting for this
target text. We add a low-level prominence to the words "You" and "it".
The word "can" is the most important word in the sentence since the
speaker is likely convincing the listener that they can do something.

# example 3
Target speaking style: Angry
Target text: "I can't stand this anymore!"

These are my suggestions:

Pitch   Energy  Duration
-1      2       -1

Word    Prominence
I       1
can't   0
stand   2
this    2
anymore 0

Explanation:
We lower the pitch and increase the energy to make the voice sound
angrier. We slightly increase the speaking rate, as this is likely
something that would be said in a heated argument. The word "I" is made
slightly more prominent since the speaker is reflecting on their own
feelings. The words "stand" and "this" are made more prominent since the
speaker probably wants to emphasise that they have had enough of
something.

# example 4
Target speaking style: A college professor giving a lecture
Target text: "The following example illustrates the point."

This is what I would suggest:

Pitch   Energy  Duration
-1       0       2

Word        Prominence
The         0
following   0
example     0
illustrates 2
the         0
point       1

Explanation:
We make the speaking rate slower since that helps the listener to follow
along. A slightly lower pitch is fitting for this target speaking style.
We only make the words "illustrates" and "point" slightly more prominent
since they are directly involved in the point the speaker is trying to
make.

# example 5
Target speaking style: A villain in a movie
Target text: "Soon, this city will be mine!"

These are my suggestions for this speaking style:

Pitch   Energy  Duration
-3      4       -2

Word    Prominence
Soon    0
```

```
this    0
city    0
will    0
be      0
mine    4

Explanation:
A lower pitch and higher energy are fitting for a voice that is
perceived as evil. We shorten the overall duration to convey excitement
in the speaker's voice. We make the word "mine" much more prominent
since it emphasises the goal of the speaker.

# example 6
Target speaking style: A child
Target text: "Can we have pizza for dinner?"

My suggestions:

Pitch   Energy  Duration
2       0       2

Word    Prominence
Can     0
we      0
have    0
pizza   2
for     0
dinner  0

Explanation:
We make the pitch higher and the speaking rate slower, which would be
appropriate for a child speaking. We emphasise the word "pizza" since
it is likely the focus of the child's request.

# example 7
Target speaking style: An old man telling a story
Target text: "In those days, we didn't have much."

I would make the following adjustments:

Pitch   Energy  Duration
-1      -3      3

Word    Prominence
In      0
those   0
days    1
we      0
didn't  0
have    0
much    0

Explanation:
The energy and speaking rate are decreased to make the voice sound older.
The pitch is slightly lowered since it is an old man. We only emphasise
the word "days" slightly since the speaker is describing a particular
time period.

# example 8
Target speaking style: A shy person
Target text: "I'm not sure what they think of me."

Here are my suggestions:

Pitch   Energy  Duration
-2      -2      2

Word    Prominence
I'm     0
not     1
```

```
sure    0
what    0
they    2
think   0
of      0
me      1

Explanation:
We decrease pitch, energy and duration, which would be appropriate for a
less confident speaker. We add a small prominence value to the word
"not" to indicate the negation. We then add prominence to the word
"they" since the speaker is shy and probably worries about what other
people think of them. We also assign a low prominence value to the word
"me" since the speaker is referring to themselves.

# example 9
Target speaking style: A child waking up on Christmas morning
Target text: "Quick, let's go and open the presents!"

I would make the following adjustments:

Pitch   Energy  Duration
2       3       -2

Word    Prominence
Quick   3
let's   0
go      0
and     0
open    0
the     0
presents 3

Explanation:
We increase pitch, energy, and speaking rate, as that would be fitting
for the excitement the child must be experiencing. We add high
prominence values to the words "Quick" and "presents". The word "Quick"
is prominent because it indicates the urgency of the situation. The word
"presents" is prominent because it explains why the situation is
exciting.

# example 10
Target speaking style: A very tired person
Target text: "We should probably go to bed."

These are the changes I would make:

Pitch   Energy  Duration
-2      -4      3

Word        Prominence
We          0
should      0
probably    0
go          0
to          0
bed         0

Explanation:
We decrease pitch, energy, and speaking rate significantly, as the
speaker is probably speaking with very low effort. We don't add any
particular prominence to this utterance, as the tone of the voice
should convey the message effectively.

Now do it for the following inputs:
```

Appendix F

A fine-tuning strategy for discovering controllable features

F.1 Instruction prompt

Listing: Instruction prompt used in Sigurgeirsson and King (2025) to generate natural language descriptions of training data

```
You will be given 7 descriptive keywords related to an audio sample
of speaker_name's speech. These keywords include:
1. The gender (male, female)
2. The level of reverberation (very distant-sounding, distant-sounding,
    slightly distant-sounding, slightly close-sounding, very close-
    sounding)
3. The amount of noise in the sample (extremely noisy, very noisy,
    noisy, slightly noisy, almost no noise, very clear)
4. The tone of the speaker's voice (very monotone, monotone, slightly
    expressive and animated, expressive and animated, very expressive
    and animated)
5. The pace of the speaker's delivery (very slowly, slowly, slightly
    slowly, moderate speed, slightly fast, fast, very fast)
6. The pitch of the speaker's voice (very low-pitch, low-pitch, slightly
    low-pitch, moderate pitch, slightly high-pitch, high-pitch, very
    high-pitch)
7. The emotion of the speaker's voice. This could be one of 6: Happy,
    Sad, Angry, Surprised, Helpful or Neutral.

This will also include the intensity of the emotion (for example:
- neutral emotion: there is no particular emotion
- high intensity sad emotion: the speaker is sad, and the intensity of
    the emotion is high
- medium intensity happy emotion: the speaker sounds happy, and the
    intensity of that emotion is medium
- low intensity surprised emotion: the speaker sounds a little bit
    surprised
- Very high intensity helpful emotion: the speaker sounds
    incredibly helpful
...)

Your task is to create a text description using these keywords that
accurately describes the speech sample. If the amount of noise is 'very
noisy' and the level of reverberation is 'very distant-sounding', you
must include terms such as 'very poor recording' or 'very bad recording'
```

in the description. Likewise, if the amount of noise is 'very clear' and the level of reverberation is 'very close-sounding', you must include terms like 'very good recording' or 'excellent recording' in the description. And you must always specify what the emotion of the speaker is, and the intensity of that emotion. For example, if the emotion of the speaker is "low intensity happy emotion", you must include terms like 'slightly happy sounding' in the description. Do not add extra details beyond what has been provided above. You can change the order of keywords and replace synonymous terms.

For example, given the following keywords: 'female', 'slightly distant-sounding', 'noisy', 'very expressive and animated', 'very slowly', 'moderate pitch' and 'high intensity angry emotion' a valid description would be: 'speaker_name speaks very slowly but has a very animated delivery. She sounds noticeably angry. The recording is noisy and there is some roominess.' Another valid description would be: 'In a noisy room, speaker_name delivers a very animated and expressive speech, at a very slow pace. speaker_name is audibly angry.' Another valid description would be: 'speaker_name enunciates a very expressive speech while clearly angry. Her voice is slightly distant-sounding, with some background noise present. speaker_name speaks very slowly with a moderate pitch but a very expressive tone.'

Note that the intensity of the speaker's emotion is sometimes specifically mentioned. This should not be confused with the speaker's tone. So, for example, the speaker might be 'very expressive and animated' and have 'low intensity sad emotion'. In which case, you have to describe the tone of voice as being expressive (e.g. 'speaker_name's tone is highly dynamic') while the intensity of the emotion is low (e.g. 'speaker_name sounds a little bit sad')

Ensure that the generated description is grammatically correct, easy to understand, and concise. Only return one and only one description.

For the keywords: 'gender', 'reverberation', 'sdr_noise', 'speech_monotony', 'speaking_rate', 'pitch' and 'emotion' the corresponding description is:

Bibliography

Masanobu Abe. Speaking styles: statistical analysis and synthesis by a Text-To-Speech system. In *Progress in Speech Synthesis*, pages 495–510. Springer, 1997.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *ArXiv Preprint*, 2023.

Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Expressive speech synthesis via modeling expressions with variational autoencoder. In *Interspeech*, 2018.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI magazine*, pages 105–120, 2014.

Sercan Ö Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep Voice: real-time neural Text-To-Speech. In *International Conference on Machine Learning (ICML)*, pages 195–204, 2017.

Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. Marta: Leveraging human rationales for explainable text classification. In *AAAI conference on artificial intelligence*, pages 5868–5876, 2021.

Amalia Arvaniti. The phonetics of prosody. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, 2020.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv Preprint*, 2016.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski,

Alexis Conneau, and Michael Auli. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech*, 2022.

Jo-Anne Bachorowski and Michael J Owren. Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, pages 219–224, 1995.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2Wec 2.0: a framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, pages 12449–12460, 2020.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ArXiv Preprint*, 2014.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, pages 1–48, 2015.

Eric Battenberg, Soroosh Mariooryad, Daisy Stanton, R. J. Skerry-Ryan, Matt Shannon, David Kao, and Tom Bagby. Effective use of variational embedding capacity in expressive end-to-end speech synthesis. *ArXiv Preprint*, 2019.

Oliver Baumann and Pascal Belin. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research*, pages 110–120, 2010.

Yoshua Bengio et al. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, pages 1–127, 2009.

Simon Betz, Birte Carlmeyer, Petra Wagner, and Britta Wrede. Interactive hesitation synthesis: modelling and evaluation. *Multimodal Technologies and Interaction*, 2018.

Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, pages 434–451, 2008.

Alan W Black, Heiga Zen, and Keiichi Tokuda. Statistical parametric speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco

Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023a.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. Soundstorm: Efficient parallel audio generation. *ArXiv Preprint*, 2023b.

Ann R Bradlow, Lynne C Nygaard, and David B Pisoni. Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, pages 206–219, 1999.

Peter D Bricker and Sandra Pruzansky. Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America*, pages 1441–1449, 1966.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, pages 1877–1901, 2020.

Kenton L Burns and Ernst G Beier. Significance of vocal and visual channels in the decoding of emotional meaning. *Journal of Communication*, pages 118–130, 1973.

Xiong Cai, Dongyang Dai, Zhiyong Wu, Xiang Li, Jingbei Li, and Helen M. Meng. Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Liping Chen, Yan Deng, Xi Wang, Frank K Soong, and Lei He. Speech Bert embedding for improving prosody in neural TTS. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

Kyunghyun Cho, B Van Merrienboer, Caglar Gulcehre, F Bougares, H Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Wei Chu and Abeer Alwan. Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 2024.

John John Ellery Clark, Colin Yallop, and Janet Fletcher. *An introduction to phonetics and phonology*. Oxford: Blackwell Publishing, 1995.

Robert AJ Clark, Korin Richmond, and Simon King. Festival 2 - build your own general purpose unit selection speech synthesiser. In *ISCA Speech Synthesis Workshop*, 2004.

William Clark, Jan Golinski, and Simon Schaffer. *The sciences in enlightened Europe*. University of Chicago Press, 1999.

William S Cleveland. Lowess: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, 1981.

Michelle Cohn and Georgia Zellou. Perception of concatenative vs. neural text-to-speech: Differences in intelligibility in noise and language attitudes. In *Interspeech*, 2020.

Jennifer Cole. Prosody in context: a review. *Language, Cognition and Neuroscience*, pages 1–31, 2015.

Jennifer Cole and Stefanie Shattuck-Hufnagel. The phonology and phonetics of perceived prosody: what do listeners imitate? In *Interspeech*, 2011.

Jennifer Cole and Stefanie Shattuck-Hufnagel. New methods for prosodic transcription: capturing variability as a source of information. *Laboratory Phonology*, 2016.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, pages 47704–47720, 2023.

Tobias Cornille, Fengna Wang, and Jessa Bekker. Interactive multi-level prosody control for expressive speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Frances S Costanzo, Norman N Markel, and Philip R Costanzo. Voice quality profile and perceived emotion. *Journal of Counseling Psychology*, 1969.

Anne Cutler and SD Isard. The production of prosody. In *Language Production*, pages 245–269. Academic Press, 1980.

Joel R Davitz and Lois Jean Davitz. The communication of feelings by content-free speech. *Journal of Communication*, 1959.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019.

N Dixon and H Maxey. Terminal analog synthesis of continuous speech using the diphone method of segment assembly. *IEEE Transactions on Audio and Electroacoustics*, 1968.

Phat Do, Matt Coler, Jelske Dijkstra, and Esther Klabbers. Text-To-Speech for under-resourced languages: Phoneme mapping and source language selection in transfer learning. In *Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 16–22, 2022.

Yadolah Dodge. *The concise encyclopedia of statistics*. Springer Science & Business Media, 2008.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *ArXiv Preprint*, 2017.

Amy Drahota, Alan Costall, and Vasudevi Reddy. The vocal communication of different kinds of smile. *Speech Communication*, pages 278–287, 2008.

Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sergiy Matusevych, Sebastian Braun, Emre Sefik Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, and Robert Aichner. ICASSP 2022 deep noise suppression challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Homer Dudley. Remaking speech. *The Journal of the Acoustical Society of America*, pages 169–177, 1939.

Homer Dudley, Richard R Riesz, and Stanley SA Watkins. A synthetic speaker. *Journal of the Franklin Institute*, pages 739–764, 1939.

Gölge Eren and The Coqui TTS Team. Coqui TTS, 2021.

Maxine Eskenazi. Trends in speaking styles research. In *European Conference on Speech Communication and Technology*, 1993.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia*, pages 1459–1462, 2010.

Gunnar Fant. Ove II synthesis strategy. In *Speech Communication Seminar*, 1952. and others.

Gunnar Fant. Acoustic theory of speech production : with calculations based on x-ray studies of russian articulations. *Slavic and East European Journal*, 1961.

Lyn Frazier, Katy Carlson, and Charles Clifton. Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences*, pages 244–249, 2006.

Robert W Frick. Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 1985.

Karl Friston, Philipp Schwartenbeck, Thomas FitzGerald, Michael Moutoussis, Timothy Behrens, and Raymond J Dolan. The anatomy of choice: active inference and agency. *Frontiers in Human Neuroscience*, 2013.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. Proceedings of Machine Learning Research (PMLR), 2015.

Rudolf P. Gaudio. Sounding gay: pitch properties in the speech of gay and straight men. *American Speech*, pages 30–57, 1994. Duke University Press, American Dialect Society.

Antonella Giannini. The two heads of the abbé. In *International Congress of Phonetic Sciences*, pages 2533–36, 1999.

Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep Voice 2: Multi-speaker neural Text-To-Speech. In *Advances in Neural Information Processing Systems*, pages 2962–2970, 2017.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana

Kagal. Explaining explanations: An overview of interpretability of machine learning. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2018.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT press Cambridge, 2016.

John D Gould and Clayton Lewis. Designing for usability: key principles and what designers think. *Communications of the ACM*, pages 300–311, 1985.

Avashna Govender and Simon King. Using pupillometry to measure the cognitive load of synthetic speech. In *Interspeech*, 2018.

Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1984.

Haohan Guo, Shaofei Zhang, Frank K Soong, Lei He, and Lei Xie. Conversational end-to-end TTS for voice agents. In *IEEE Spoken Language Technology Workshop (SLT)*, 2021.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. PromptTTS: Controllable Text-To-Speech with text descriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Joakim Gustafson, Jonas Beskow, and Eva Székely. Personality in the mix-investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis. In *ISCA Speech Synthesis Workshop*, 2021.

Peter Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. Gibbs sampling with people. *Advances in Neural Information Processing Systems*, pages 10659–10671, 2020.

Valerie Hazan and Rachel Baker. Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style? In *Workshop on Disfluency in Spontaneous Speech and the 2nd International Symposium on Linguistic Patterns in Spontaneous Speech*, pages 7–10, 2010.

Elina Helander and Jani Nurminen. On the importance of pure prosody in the perception of speaker identity. In *Interspeech*, pages 2665–2668, 2007.

Gustav Eje Henter, Jaime Lorenzo-Trueba, Xin Wang, Mariko Kondo, and Junichi Yamagishi. Cyborg speech: Deep multilingual speech synthesis for generating segmental foreign accent with natural prosody. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018a.

Gustav Eje Henter, Xin Wang, and Junichi Yamagishi. Deep encoder-decoder models for unsupervised learning of controllable speech synthesis. *ArXiv Preprint*, 2018b.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Conference on Speech and Computer*, pages 198–208. Springer, 2018.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv Preprint*, 2012.

Julia Hirschberg. A corpus-based approach to the study of speaking style. In *Prosody: Theory and Experiment*, pages 335–350. Springer, 2000.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020.

Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, pages 107–116, 1998.

Zack Hodari, Oliver Watts, and Simon King. Using generative modelling to produce varied intonation for speech synthesis. In *ISCA Speech Synthesis Workshop*, pages 239–244, 2019.

Rüdiger Hoffmann. The long way of speech synthesis. In *International Workshop on the History of Speech Communication Research*, 2019.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979.

John N Holmes, Ignatius G Mattingly, and John N Shearme. Speech synthesis by rule. *Language and Speech*, pages 127–143, 1964.

Homer and Augustus Taber Murray. *The Iliad with an English translation by AT Murray*. London; GP Putnam's Sons: New York, 1924.

Wei-Ning Hsu, Yu Zhang, Ron Weiss, Heiga Zen, Yonghui Wu, Yuan Cao, and Yuxuan Wang. Hierarchical generative modeling for controllable speech synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu. InstructTTSEval: Benchmarking complex natural-language instruction following in Text-To-Speech systems. *ArXiv Preprint*, 2025.

Keith Ito and Linda Johnson. The LJ Speech dataset, 2017.

Je Hun Jeon and Yang Liu. Automatic prosodic events detection using syllable-based acoustic and syntactic features. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. Textrolspeech: A text style control speech corpus with codec language Text-To-Speech models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al. Transfer learning from speaker verification to multispeaker Text-To-Speech synthesis. In *International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4485–4495, 2018.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *ArXiv Preprint*, 2023.

Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. Speechcraft: A fine-grained expressive speech dataset with nat-

ural language description. In *ACM International Conference on Multimedia*, pages 1255–1264, 2024.

Patrik N Juslin and Petri Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 2003.

Sven Kachel, Manuel Pöhlmann, and Christine Nussbaum. Queer Events, Relationships, and Sports: Does Topic Influence Speakers' Acoustic Expression of Sexual Orientation? In *Interspeech*, 2023.

Rudolf E Kalman. On the general theory of control systems. In *International Conference on Automatic Control*, pages 481–492, 1960.

Sotirios Karapiperis, Nikolaos Ellinas, Alexandra Vioni, Junkwang Oh, Gunu Jho, Inchul Hwang, and Spyros Raptis. Investigating disentanglement in a phoneme-level speech codec for prosody modeling. In *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

Sri Karlapati, Alexis Moinet, Arnaud Joly, Viacheslav Klimkov, Daniel Sáez-Trigueros, and Thomas Drugman. CopyCat: many-to-many fine-grained prosody transfer for neural Text-To-Speech. In *Interspeech*, 2020.

Sri Karlapati, Penny Karanasou, Mateusz Lajszczak, Syed Ammar Abbas, Alexis Moinet, Peter Makarov, Ray Li, Arent van Korlaar, Simon Slangen, and Thomas Drugman. CopyCat2: a single model for multi-speaker TTS and many-to-many fine-grained prosody transfer. In *Interspeech*, 2022.

Hideki Kawahara. STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, pages 349–353, 2006.

Tom Kenter, Vincent Wan, Chun-An Chan, Rob Clark, and Jakub Vit. Chive: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network. In *International Conference on Machine Learning*, pages 3331–3340. Proceedings of Machine Learning Research (PMLR), 2019.

Peter Kieseberg, Edgar Weippl, A Min Tjoa, Federico Cabitza, Andrea Campagner, and Andreas Holzinger. Controllable AI–an alternative to trustworthiness in complex AI systems? In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2023.

Simon King. A tutorial on HMM speech synthesis. *Sadhana: Academy Proceedings in Engineering Sciences*, 2010.

Simon King and Vasilis Karaiskos. The blizzard challenge 2013. In *The Blizzard Challenge 2013*, pages 1–12, 2013.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ArXiv Preprint*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ArXiv Preprint*, 2013.

Dennis H Klatt. Review of Text-To-Speech conversion for english. *The Journal of the Acoustical Society of America*, pages 737–793, 1987.

Viacheslav Klimkov, Srikanth Ronanki, Jonas Rohnke, and Thomas Drugman. Fine-grained robust prosody transfer for single-speaker neural Text-To-Speech. In *Interspeech*, 2019.

Daichi Kondo and Masanori Morise. Human-In-The-Loop speech-design system and its evaluation. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 608–612, 2019.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-Gan: generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, pages 17022–17033, 2020.

Christian Gottlieb Kratzenstein. *Tentamen resolvendi problema*. TUDpress, 1781.

Jody Kreiman and Diana Sidtis. *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons, 2011.

Jody Kreiman, Bruce R Gerratt, Kristin Precoda, and Gerald S Berke. Individual differences in voice quality perception. *Journal of Speech, Language, and Hearing Research*, pages 512–520, 1992.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, pages 583–621, 1952.

Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1993.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2018.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGan. *Advances in Neural Information Processing Systems*, pages 27980–27993, 2023.

Ohsung Kwon, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang. An effective style token weight control technique for end-to-end emotional speech synthesis. *IEEE Signal Processing Letters*, 2019.

Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. Parler-TTS, 2024. Available at: https://github.com/huggingface/parler-tts. Accessed: April 2025.

D Robert Ladd. *Intonational phonology*. Cambridge University Press, 2008.

Adrian Lańcucki. Fastpitch: Parallel Text-To-Speech with pitch prediction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

Nadine Lavan, A Mike Burton, Sophie K Scott, and Carolyn McGettigan. Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, pages 90–102, 2019.

Walter Lawrence. The synthesis of speech from signal which have a low information rate. *Communication Theory*, pages 460–469, 1953.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. SDR – half-baked or well done? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, pages 436–444, 2015.

Younggun Lee and Taesu Kim. Robust and fine-grained prosody control of end-to-end speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

Yichong Leng, Zhifang Guo, Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, Lei He, Xiangyang Li, Sheng Zhao,

Tao Qin, and Jiang Bian. PromptTTS 2: Describing and generating voices with text prompt. In *International Conference on Learning Representations (ICLR)*, 2024.

Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: building a better stroke prediction model. *Annals of Applied Statistics*, 2015.

Erez Levon. Sexuality in context: variation and the sociolinguistic perception of identity. *Language in Society*, pages 533–554, 2007.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *AAAI conference on artificial intelligence*, 2019.

Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G Ward. On the utility of self-supervised models for prosody-related tasks. In *IEEE Spoken Language Technology Workshop (SLT)*, 2023.

Guanghou Liu, Yongmao Zhang, Yinjiao Lei, Yunlin Chen, Rui Wang, Zhifei Li, and Linfu Xie. PromptStyle: Controllable style transfer for Text-To-Speech with natural language descriptions. In *Interspeech*, 2023.

Songxiang Liu, Disong Wang, Yuewen Cao, Lifa Sun, Xixin Wu, Shiyin Kang, Zhiyong Wu, Xunying Liu, Dan Su, Dong Yu, et al. End-to-end accent conversion without using native utterances. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019.

Zhe Liu, Yufan Guo, and Jalal Mahmud. When and why a model fails? a Human-In-The-Loop error detection framework for sentiment analysis. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 170–177, 2021.

Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 1982.

Dan Lyth and Simon King. Natural language guidance of high-fidelity Text-To-Speech with synthetic annotations, 2024.

Sara Mack and Benjamin Munson. The influence of /s/ quality on ratings of men's sexual orientation: explicit and implicit measures of the 'gay lisp' stereotype. *Journal of Phonetics*, pages 198–212, 2012.

Peter Makarov, Syed Ammar Abbas, Mateusz Lajszczak, Arnaud Joly, Sri Karlapati, Alexis Moinet, Thomas Drugman, and Penny Karanasou. Simple and effective multi-sentence TTS with expressive and coherent prosody. In *Interspeech*, 2022.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60, 1947.

Himanshu Maurya and Atli Thor Sigurgeirsson. A Human-In-The-Loop approach to improving cross-text prosody transfer. In *Interspeech*, 2024.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, 2017.

Patrick E McKight and Julius Najab. Kruskal-wallis test. *The Corsini Encyclopedia of Psychology*, 2010.

Larry R Medsker, Lakhmi Jain, et al. Recurrent neural networks. *Design and Applications*, page 2, 2001.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, pages 1–38, 2019.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022.

Devang S Ram Mohan, Vivian J Hu, Tian Huey Teh, Alexandra Torresquintero, Christopher GR Wallis, Marlene Staib, Lorenzo Foglianti, Jiameng Gao, and Simon King. Ctrl-p: Temporal control of prosodic variation for speech synthesis. In *Interspeech*, 2021.

James N Morgan and John A Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, pages 415–434, 1963.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, pages 1877–1884, 2016.

Max Morrison, Caedon Hsieh, Nathan Pruyne, and Bryan Pardo. Cross-domain neural pitch and periodicity estimation. *ArXiv Preprint*, 2023.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-In-The-Loop machine learning: a state of the art. *Artificial Intelligence Review*, pages 3005–3054, 2023.

Anthony Mulac and Howard Giles. 'your're only as old as you sound': Perceived vocal age and social meanings. *Health Communication*, pages 199–215, 1996.

Benjamin Munson. Lavender Lessons Learned; Or, What Sexuality Can Teach Us About Phonetic Variation. *American Speech*, pages 14–31, 2011.

Benjamin Munson, Sarah V. Jefferson, and Elizabeth C. McDonald. The influence of perceived sexual orientation on fricative identification. *The Journal of the Acoustical Society of America*, pages 2427–2437, 2006.

Iain R Murray and John L Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, pages 1097–1108, 1993.

Tomohiro Nakatani, Shigeaki Amano, Toshio Irino, Kentaro Ishizuka, and Tadahisa Kondo. A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments. *Speech Communication*, pages 203–214, 2008.

Anna Björk Nikulásdóttir, Jón Guðnason, and Eiríkur Rögnvaldsson. An icelandic pronunciation dictionary for TTS. In *IEEE Spoken Language Technology Workshop (SLT)*, 2018.

John J Ohala. Review of suprasegmentals. *Language*, pages 736–740, 1975.

Pilar Oplustil-Gallegos and Simon King. Using previous acoustic context to improve Text-To-Speech synthesis. *ArXiv Preprint*, 2020.

Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TAdam: Task dependent adaptive metric for improved few-shot learning. *Advances in Neural Information Processing Systems*, 2018.

Gunnar Thor Örnólfsson, Atli Thor Sigurgeirsson, Anna Björk Nikulásdóttir, and Daniel Schnell. Talrómur 3 v0.1 (24.09), 2024. Clarin-IS.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Stephen A Paipetis. *Science and Technology in Homeric Epics*. Springer Science & Business Media, 2008.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

Nishant Prateek, Mateusz Lajszczak, Roberto Barra-Chicote, Thomas Drugman, Jaime Lorenzo-Trueba, Thomas Merritt, Srikanth Ronanki, and Trevor Wood. In other news: a bi-style Text-To-Speech model for synthesizing newscaster voice with limited data. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 205–213, 2019.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. Proceedings of Machine Learning Research (PMLR), 2019.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pages 7836–7846. Proceedings of Machine Learning Research (PMLR), 2020.

Tuomo Raitio, Ramya Rasipuram, and Dan Castellani. Controllable neural Text-To-Speech synthesis using intuitive prosodic features. In *Interspeech*, 2020.

Tuomo Raitio, Jiangchuan Li, and Shreyas Seshadri. Hierarchical prosody modeling and control in non-autoregressive parallel neural TTS. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022a.

Tuomo Raitio, Petko N. Petkov, Jiangchuan Li, Muhammed P.V. Shifas, Andrea Davis, and Yannis Stylianou. Vocal effort modeling in neural TTS for improving the intelligibility of synthetic speech in noise. In *Interspeech*, 2022b.

Tamara Rakić, Melanie C Steffens, and Amélie Mummendey. When it matters how you pronounce it: The influence of regional accents on job interview outcome. *British Journal of Psychology*, pages 868–883, 2011.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. Proceedings of Machine Learning Research (PMLR), 2021.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 2019.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations (ICLR)*, 2020.

Manuel Sam Ribeiro et al. Parallel audiobook corpus, 2018.

Albert Rilliard, Alexandre Allauzen, and Philippe Boula_de_Mareüil. Using dynamic time warping to compute prosodic similarity measures. In *Annual Conference of the International Speech Communication Association (ISCA)*, 2011.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual

evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.

Peter Roach. *English phonetics and phonology*. Cambridge University Press, 1989.

Andrew Ross, Nina Chen, Elisa Zhao Hang, Elena L Glassman, and Finale Doshi-Velez. Evaluating the interpretability of generative models by interactive reconstruction. In *Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California University, 1985.

Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Perceptual-similarity-aware deep speaker representation learning for multi-speaker generative modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

Adam Sanborn and Thomas Griffiths. Markov chain monte carlo with people. *Advances in Neural Information Processing Systems*, 2007.

Klaus R Scherer, Tom Johnstone, and Gundrun Klasmeyer. Vocal expression of emotion. *Handbook of Affective Sciences*, pages 433–456, 2003.

Stefan R Schweinberger, Hideki Kawahara, Adrian P Simpson, Verena G Skuk, and Romi Zäske. Speaker perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, pages 15–25, 2014.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *Python in Science*, 2010.

Shreyas Seshadri, Tuomo Raitio, Dan Castellani, and Jiangchuan Li. Emphasis control for parallel neural TTS. In *Interspeech*, 2022.

Burr Settles. Active learning literature survey. *Science*, pages 237–304, 1995.

Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality. *Biometrika*, pages 591–611, 1965.

Slava Shechtman, Raul Fernandez, and David Haws. Supervised and unsupervised approaches for controlling narrow lexical focus in sequence-to-sequence speech synthesis. In *IEEE Spoken Language Technology Workshop (SLT)*, 2021.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural TTS synthesis by conditioning Wavenet on mel spectrogram predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Jonathan Shen, Ye Jia, Mike Chrzanowski, Yu Zhang, Isaac Elias, Heiga Zen, and Yonghui Wu. Non-attentive Tacotron: Robust and controllable neural TTS synthesis including unsupervised duration modeling. *ArXiv Preprint*, 2020.

Atli Thor Sigurgeirsson and Simon King. Do prosody transfer models transfer prosody? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Atli Thor Sigurgeirsson and Simon King. Controllable speaking styles using a large language model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Atli Thor Sigurgeirsson and Simon King. RepeaTTS: Towards feature discovery through repeated fine-tuning, 2025.

Atli Thor Sigurgeirsson and Eddie L Ungless. Just because we camp, doesn't mean we should: The ethics of modelling queer voices. In *Interspeech*, 2024.

Atli Thor Sigurgeirsson, Þorsteinn Daði Gunnarsson, Gunnar Thor Örnólfsson, Eydís Huld Magnúsdottir, Kr. Þórhallsdóttir, Ragnheiður, Stefán Gunnlaugur Jónsson, and Jón Guðnason. Talrómur: A large Icelandic TTS corpus, 2021.

Henrik Singmann, Ben Bolker, Jake Westfall, Frederik Aust, and Mattan S. Ben-Shachar. *Afex: analysis of Factorial Experiments*, 2024. R package version 1.4-1.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In *International conference on machine learning*. Proceedings of Machine Learning Research (PMLR), 2018.

Ron Smyth, Greg Jacobs, and Henry Rogers. Male voices and perceived sexual orientation: An experimental and theoretical approach. *Language in Society*, pages 329–350, 2003.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output represen-

tation using deep conditional generative models. *Advances in Neural Information Processing Systems*, 2015.

Alexander Sorin, Slava Shechtman, and Ron Hoory. Principal style components: Expressive style control and cross-speaker transfer in neural TTS. In *Interspeech*, 2020.

Jose M. R. Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, Aaron C. Courville, and Yoshua Bengio. Char2Wav: end-to-end speech synthesis. In *International Conference on Learning Representations (ICLR)*, 2017.

Richard Sproat and Navdeep Jaitly. RNN approaches to text normalization: A challenge. *ArXiv Preprint*, 2016.

Steinþór Steingrímsson, Iben Nyholm Debess, Kimmo Granqvist, Per Langgård, and Trond Trosterud. *Language Technology for Less-Resourced Languages in the Nordics: Current Developments and Collaborative Opportunities*. Stjórnarráð Íslands, 2024.

Kenneth N Stevens, Stanley Kasowski, and C Gunnar M Fant. An electrical analog of the vocal tract. *The Journal of the Acoustical Society of America*, pages 734–742, 1953.

John Q Stewart. An electrical analogue of the vocal organs. *Nature*, pages 311–312, 1922.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, pages 3008–3021, 2020.

Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

Guangzhi Sun, Yu Zhang, Ron J Weiss, Yuan Cao, Heiga Zen, and Yonghui Wu. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 2014.

Éva Székely, Jūra Miniota, et al. Will AI shape the way we speak? the emerging sociolinguistic influence of synthetic voices. *ArXiv Preprint*, 2025.

Fabio Tamburini. Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system. In *European Conference on Speech Communication and Technology*, 2003.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *ArXiv Preprint*, 2021.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. NaturalSpeech: End-to-end Text-To-Speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Vivien C Tartter and David Braun. Hearing smiles and frowns in normal and whisper registers. *The Journal of the Acoustical Society of America*, pages 2101–2107, 1994.

Jason Taylor and Korin Richmond. Analysis of pronunciation learning in end-to-end speech synthesis. In *Interspeech*, 2019.

Paul Taylor. *Text-To-Speech synthesis*. Cambridge university press, 2009.

Ryunen Teranishi and Noriko Umeda. Use of pronouncing dictionary in speech synthesis experiments. In *International Congress on Acoustics*, pages 155–158, 1968.

Noé Tits, Kevin El Haddad, and Thierry Dutoit. Ice-Talk 2: Interface for controllable expressive TTS with perceptual assessment tool. *Software Impacts*, 2021.

Tomoki Toda, Alan W Black, and Keiichi Tokuda. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

Keiichi Tokuda, Heiga Zen, and Alan W Black. An HMM-based speech synthesis system applied to english. In *IEEE Speech Synthesis Workshop (SSW)*, pages 227–230, 2002.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *ArXiv Preprint*, 2023.

Kenta Udagawa, Yuki Saito, and Hiroshi Saruwatari. Human-In-The-Loop speaker adaptation for DNN-based multi-speaker tts. In *Interspeech*, 2022.

Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. MelloTron: multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020a.

Rafael Valle, Kevin J Shih, Ryan Prenger, and Bryan Catanzaro. FlowTron: an autoregressive flow-based generative network for Text-To-Speech synthesis. In *International Conference on Learning Representations (ICLR)*, 2020b.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *ArXiv Preprint*, 2016.

Pol van Rijn, Silvan Mertes, Dominik Schiller, Peter M C Harrison, Pauline Larrouy-Maestri, Elisabeth Andre, and Nori Jacoby. Exploring emotional prototypes in a high dimensional TTS latent space. In *Interspeech*, 2021.

Jan PH Van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg. *Progress in speech synthesis*. Springer Science & Business Media, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Wolfgang Von Kempelen. *Mechanismus der menschlichen Sprache*. Degen, 1791.

Jovana Vukovic, Benedict C Jones, David R Feinberg, Lisa M DeBruine, Finlay G Smith, Lisa LM Welling, and Anthony C Little. Variation in perceptions of physical dominance and trustworthiness predicts individual differences in the effect of relationship context on women's preferences for masculine pitch in men's voices. *British Journal of Psychology*, pages 37–48, 2011.

Michael Wagner and Duane G Watson. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, pages 905–945, 2010.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *ArXiv Preprint*, 2023.

Wenfu Wang, Shuang Xu, Bo Xu, et al. First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention. In *Interspeech*, 2016.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: towards end-to-end speech synthesis. *Interspeech*, 2017.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pages 5180–5189. Proceedings of Machine Learning Research (PMLR), 2018.

A Watanabe, S Felicetti, B Hedström, G Surjadi, G Tannergård, I Tegerstedt, B Wejnebring, M-B Wetterling, L Andersson, L Hallsten, and et al. Gunnar fant 60 years. *TMH-QPSR*, page 1–45, 1979.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Thomas W. Williams. Our exhibits at two fairs. *Bell System Technical Journal*, 1940.

Deirdre Wilson and Tim Wharton. Relevance and prosody. *Journal of Pragmatics*, pages 1559–1579, 2006.

Werner Wirth and Holger Schramm. Media and emotions. *Communication Research Trends*, 2005.

Pengfei Wu, Zhenhua Ling, Lijuan Liu, Yuan Jiang, Hongchuan Wu, and Lirong Dai. End-to-end emotional speech synthesis using style tokens and semi-supervised training. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A

survey of Human-In-The-Loop for machine learning. *Future Generation Computer Systems*, pages 364–381, 2022.

Yujia Xiao, Lei He, Huaiping Ming, and Frank K Soong. Improving prosody with linguistic and Bert derived features in multi-speaker based mandarin chinese neural TTS. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Detai Xin, Yuki Saito, Shinnosuke Takamichi, Tomoki Koriyama, and Hiroshi Saruwatari. Cross-lingual Text-To-Speech synthesis via domain adaptation and perceptual similarity regression in speaker space. In *Interspeech*, 2020.

Junichi Yamagishi, Koji Onishi, Takashi Masuko, and Takao Kobayashi. Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, pages 502–509, 2005.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. InstructTTS: Modelling expressive TTS in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Kaisheng Yao and Geoffrey Zweig. Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. *Interspeech*, 2015.

Hyun-Wook Yoon, Ohsung Kwon, Hoyeon Lee, Ryuichi Yamamoto, Eunwoo Song, Jae-Min Kim, and Min-Jae Hwang. Language model-based emotion prediction methods for emotional speech synthesis systems. In *Interspeech*, 2022.

Julian Zaïdi, Hugo Seuté, Benjamin van Niekerk, and Marc-André Carbonneau. Daft-Exprt: cross-speaker prosody transfer on any text for expressive speech synthesis. In *Interspeech*, 2022.

Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong

Wu, and Jia Jia. VoxInstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. In *ACM International Conference on Multimedia*, pages 554–563, 2024.