Modeling vocal effort in speech synthesis with variational autoencoders

Daniel Lyth

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2021

Abstract

Recent work combining neural text-to-speech and probabilistic latent variable models has demonstrated efficacy in modeling the rich non-lexical information found in speech. Variables such as pitch, energy, and duration are typically investigated. However, one critical attribute of speech that is under-explored in this context is vocal effort. Using a novel dataset containing the full spectrum of vocal effort, we propose a variational autoencoder text-to-speech system capable of modeling this attribute. We explore the considerations of training such a model and discuss the appropriate level of disentanglement for this task. We go on to demonstrate the ability to synthesize speech across a wide range of vocal effort in a reliable and interpretable manner, smoothly interpolating across this distribution.

Acknowledgments

Thanks to Jason Fong, Jonathan Dyke and Korin Richmond for their numerous insights and suggestions. Thanks to Alastair MacGregor for his support. Thanks to my wife and family for their patience and encouragement.

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.

(*Daniel Lyth*)

Table of Contents

Chapter 1

Introduction

Speech synthesis is an essential component in human-computer interaction, assistive technologies in healthcare, and increasingly, in media and entertainment. Human speech communication relies on a rich flow of non-lexical information [1], and so for speech synthesis to be effective in these domains, it must do more than simply recreate an intelligible sequence of words. Indeed, text-to-speech (TTS) [2] can be considered a one-to-many mapping problem, as there are many ways in which a given sequence of text may be reasonably realized. This non-lexical variation in speech includes rhythm, stress, and intonation and is broadly referred to as prosody [3, 4]. There is no formal list of prosodic variables, but typically, acoustic features such as pitch, energy, and phone duration are considered significant contributors.

One variable that is discussed less frequently in the context of TTS is vocal effort. This attribute is most commonly defined as the way in which a speaker will modify their speech as the speaker-listener distance changes [5, 6]. However, vocal effort is also correlated to the level of background noise [7, 8], room acoustics [9, 10], and emotional state [11, 12]. Whispering and shouting are examples of low and high vocal effort, respectively. Vocal effort is a significant source of non-lexical variability in speech, and yet there has been limited TTS research in recent years that attempts to model this attribute in a probabilistic manner. We wish to address this paucity.

In the past, prosodic variation has been predominantly achieved by modeling pitch and duration and using these values to drive unit selection in concatenative synthesis [13] or vocoder parameters in statistical parametric synthesis [14]. However, the generated speech from both of these approaches suffers from a lack of naturalness. By contrast, the advent of more modern neural network-based TTS systems has brought a significant increase in naturalness [15–20]. Unfortunately, early attempts came at the cost of less prosodic control, as no mechanism was provided to model this information. Equally problematic is that the optimization of these models has an averaging effect on prosody [21]. One approach to alleviating this issue is to combine a neural TTS model with a probabilistic latent variable model, such as a variational autoencoder (VAE) [22, 23]. Ideally, this model will learn a distribution of the residual non-lexical information in the latent space from which we can sample at inference.

There have been several successful demonstrations of this approach [24–29]. However, these works have focused on modeling attributes such as pitch, energy, and speaking rate [29], speaker identity and accent [24], and emotion [26]. A critical attribute of speech that has not been explored in this context is vocal effort. This leads us to our central question – can we model vocal effort in a latent variable TTS system?

One of the reasons vocal effort is an under-explored attribute in speech synthesis is the lack of appropriate data. Typical TTS datasets are either constructed from existing audiobook recordings [30–32], recorded to create neutral-sounding voice assistants or designed for speaker adaptation or voice conversion research [33]. More naturalistic datasets that have the potential to contain a wider range of vocal effort do exist. These are typically gathered 'in-the-wild' from podcasts [34], videos [35], and other media [36, 37]. However, these datasets tend to be challenging, given the lack of single speaker data, the wildly varying channel conditions, and the lack of transcriptions. There have been a small number of datasets recorded specifically for exploring vocal effort. However, these tend to be limited in size and focus on a few discrete classes of vocal effort rather than the entire spectrum [38–44].

For this work, we use a novel dataset containing over seventeen hours of highly expressive and naturalistic speech with a wide range of vocal effort. We use this dataset in conjunction with a probabilistic latent variable TTS system in order to model the vocal effort distribution and sample from it at inference. This approach contrasts with the existing research that tends to model vocal effort with only two or three discrete classes [38–44] and provides no probabilistic method for smoothly sampling from this distribution. Prior work also typically focuses on the low-end of the vocal effort spectrum (whispered, normal, and Lombard speech) with the occasional exception [45].

Our core contributions are as follows:

- We demonstrate that a VAE-TTS system is capable of modeling vocal effort. Through quantitative analysis and subjective listening tests, we show that we can synthesize speech across a wide range of vocal effort in a reliable and interpretable manner, smoothly interpolating across this distribution. To the best of our knowledge, this is the first time that this speech attribute has been modeled in this manner.

- We provide insights into architectural and optimization considerations when training such a VAE-TTS model. These include the impact of the latent space dimensionality on disentanglement and methods to avoid the collapse of the latent space.
- We demonstrate that it is possible to train a TTS model using a dataset containing far more stylistic and vocal effort variation than typical datasets, albeit with slightly less subjective naturalness.
- We show initial promising results that this model can also capture vocal effort from a reference utterance and transfer this to a synthesized utterance.

The remainder of this dissertation is laid out as follows:

- Background and Related Work

  A brief overview of the literature regarding vocal effort, TTS, and variational autoencoders.

- Dataset

  A description of the dataset we use in this project, how we prepared it, and a comparison with a more typical TTS dataset.

- Method

  A description of the approach taken in this work, including the models used, the specific questions we choose to explore, and our criteria for success.

- Experiments and Results

  Details of the experiments we ran and corresponding results.

- Discussion

  An in-depth evaluation of our results and discussion in the context of the broader literature.

- Conclusion

  Closing remarks and a discussion of the limitations of our work and potential areas for future work.

Chapter 2

Background and Related Work

2.1 Vocal effort

Vocal effort is dependent on the speaker-listener distance [5, 6], the level of background noise (referred to as the Lombard effect) [7, 8], room acoustics [9, 10], and emotional state [11, 12]. It is typically measured in physiological terms due to the corresponding changes that take place in the voice production system [46, 47] (although some argue that it should be considered a subjective psychological measure [48]).

Vocal effort also corresponds to a variety of acoustic measures. Sound pressure level is perhaps the most notable, but several other variables such as fundamental frequency (F_0), phone duration, and spectral tilt are affected by changes in vocal effort. A comprehensive review of the correlation between perceived vocal effort and these acoustic measures can be found in [5].

In relation to emotion, vocal effort is most closely correlated to arousal [11]. For example, high arousal states such as ‘angry’ and ‘excited’ correlate to high vocal effort, and low arousal states such as ‘bored’ or ‘calm’ correlate to low vocal effort.

In the TTS research literature, vocal effort is somewhat under-explored. There have been attempts to model vocal effort in concatenative synthesis [38, 39], statistical parametric synthesis [40, 41, 45], and more recently in neural systems [42–44]. However, as mentioned in Chapter 1, these works have typically modeled only a few discrete classes of vocal effort and have tended to focus on the lower end of the vocal effort continuum. These classes are typically whispered, ‘normal’ and Lombard speech (the Lombard effect describes the involuntary increase in vocal effort due to background noise [7, 8]). However, there are occasional exceptions; [45] investigates shouted speech, for example. There have also been efforts to model vocal effort in voice conversion [49–51], automatic speech recognition (ASR) [44, 52], and speaker verification [53], although again, this is relatively under-explored.

2.2 Text-to-speech synthesis

A variety of approaches to speech synthesis have been developed over the years. These include articulatory synthesis [54, 55], formant synthesis [56, 57], concatenative synthesis [58, 59], and statistical parametric synthesis [60, 61]. However, since the advent of deep learning, artificial neural network-based TTS has become the dominant approach [15–20]. These models typically tackle the sequence-to-sequence nature of TTS using encoder-decoder architectures, inspired by those first proposed in machine translation [62, 63]. While sometimes claiming to be ‘end-to-end’, these models typically convert text into a spectrogram that must then be vocoded into a waveform using a separate model (a spectrogram is a two-dimensional representation of audio created using the short-term Fourier transform that can be visualized as an image). These vocoders are also typically neural networks, and a variety of approaches have been employed. These include the use of autoregressive [64], adversarial [65], flow-based [66], and diffusion [67] models.

The naturalness and intelligibility of neural TTS marks significant progress over previous approaches. With sufficient training data, it has been demonstrated that artificial speech can be almost imperceptible to natural speech [16, 20]. However, it can be argued that the ‘black-box’ nature of deep neural networks makes these models less controllable and interpretable than previous approaches. With early neural TTS models, controlling even relatively simple attributes such as speaking rate was not possible. Since then, significant effort has gone into regaining control mechanisms, and various approaches to modeling the non-lexical variation in speech have been proposed [19, 24, 68, 69]. Among the most promising is the use of latent variable models such as the variational autoencoder [22, 23].

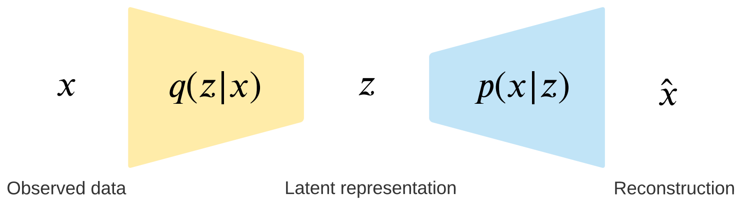


Figure 2.1: Visualization of a variational autoencoder.

2.3 Variational autoencoder

Generative models assume an underlying probability distribution is responsible for the observed variation in the data and attempt to model this distribution. One such generative model is the variational autoencoder (VAE) [22, 23]. Similar to regular autoencoders, VAEs rely on a ‘bottleneck’ between the encoder and decoder to create a lower-dimensional latent (or hidden) representation z in an unsupervised manner (see Figure 2.1). However, instead of encoding data to individual unrelated points in this latent space, VAEs use variational inference to model probability distributions.

We assume that a hidden variable z is responsible for the variation found in the observable data x. Therefore, we wish to model $p(z|x)$:

$$p(z \mid x) = \frac{p(x \mid z)p(z)}{p(x)} \tag{2.1}$$

However, computing the $p(x)$ term is intractable. To circumvent this issue, we use variational inference to provide an estimate. First, we approximate the true distribution $p(z|x)$ with a learned distribution $q(z|x)$. Typically we assume that p is an isotropic zero-mean Gaussian. Then, to ensure that the learned distribution q is close to the true distribution p, we use the Kullback–Leibler (KL) divergence to minimize the differences between these distributions:

$$\min KL(q(z \mid x) \| p(z \mid x)) \tag{2.2}$$

This KL divergence loss ensures a well-regularized latent space with closely matched p and q distributions. However, we also wish to recreate approximations $\hat{x}$ that are as similar as possible to x. To achieve this, we introduce a second loss term, the reconstruction loss. Together, the full loss is:

$$\mathcal{L}(x,\hat{x}) + \sum_j KL\left(q_j(z \mid x) \| p(z)\right) \tag{2.3}$$

where the reconstruction loss is the first term, the KL divergence is the second term, and j is the dimension of the latent space. Further proof is provided in [22].

Initial demonstrations of variational autoencoders demonstrated their ability to disentangle, or factorize, the axes of variation found in x to separate dimensions in the latent space z without supervision [70]. Furthermore, the variational objective ensures that these representations are well-formed and smoothly distributed.

2.4 VAE-TTS

Combining a VAE with a TTS model has been shown to be an effective approach to modeling the non-lexical variation in speech [24–28, 71, 72]. Several variations on this approach have been proposed. [24] uses a hierarchy of VAEs in a Gaussian mixture model to model categorical and continuous variables separately. [26] proposes a method that contrasts the typical unsupervised method of training VAEs with the use of labeled data and a mechanism for semi-supervision. [72] explores the use of a learned prior as opposed to the typical standard Gaussian. [27] proposes a method for modifying the variational posterior to be closer to the true posterior and [29] demonstrates a hierarchical conditional VAE that operates at the utterance, word and phone domains. However, none of these works have attempted to model vocal effort.

One key advantage of VAE-TTS systems is the ability to sample from a distribution of disentangled stylistic variables. This allows for smooth interpolation across these variables in an interpretable manner. By contrast, this is not possible with simple reference encoders [73] or ‘Global Style Tokens’ [74] since they do not model probability distributions.

A final consideration with VAE-TTS systems is the desired domain of disentanglement. Some work claims to achieve very effective disentanglement of low-level acoustic attributes such as F_0, energy, phone duration, or speaking rate [29]. However, semantically meaningful measures of speech such as affect and vocal effort are highly correlated to these acoustic variables [5]. We argue, therefore, that for the sake of creative control, we should focus on disentangling this semantically meaningful variation as opposed to granular acoustic features (which we accept may remain entangled).

Chapter 3

Dataset

3.1 Description

We use a proprietary American English single-speaker dataset created as part of a media entertainment project. This dataset contains a wide range of expressive and naturalistic speaking styles, and the actor was directed to perform in a realistic manner for a wide variety of life-like scenarios. The resulting corpus contains a full range of emotion, vocal effort, sentence length, and speaking rates and also contains many non-word sounds such as laughter, breaths, and grunts. This contrasts sharply with typical speech datasets used for TTS. Figure 3.1 compares our dataset to the LJ Speech dataset [75] (an audiobook dataset commonly used in TTS research) across a variety of utterance-level statistics. One of the key challenges of our work is using a dataset with this amount of variation for TTS.

3.2 Preparation

3.2.1 Cleaning

To prepare the dataset, we first removed utterances containing no words and only non-word vocalizations such as breaths, grunts, or humming. We then trimmed any sections of silence longer than 300ms. This was done to help the sequence-to-sequence part of the model learn alignment while still allowing a reasonable amount of naturalness in speaking rhythm and pauses. The resulting clean dataset contains 25,829 utterances and is 17.2 hours long in total.

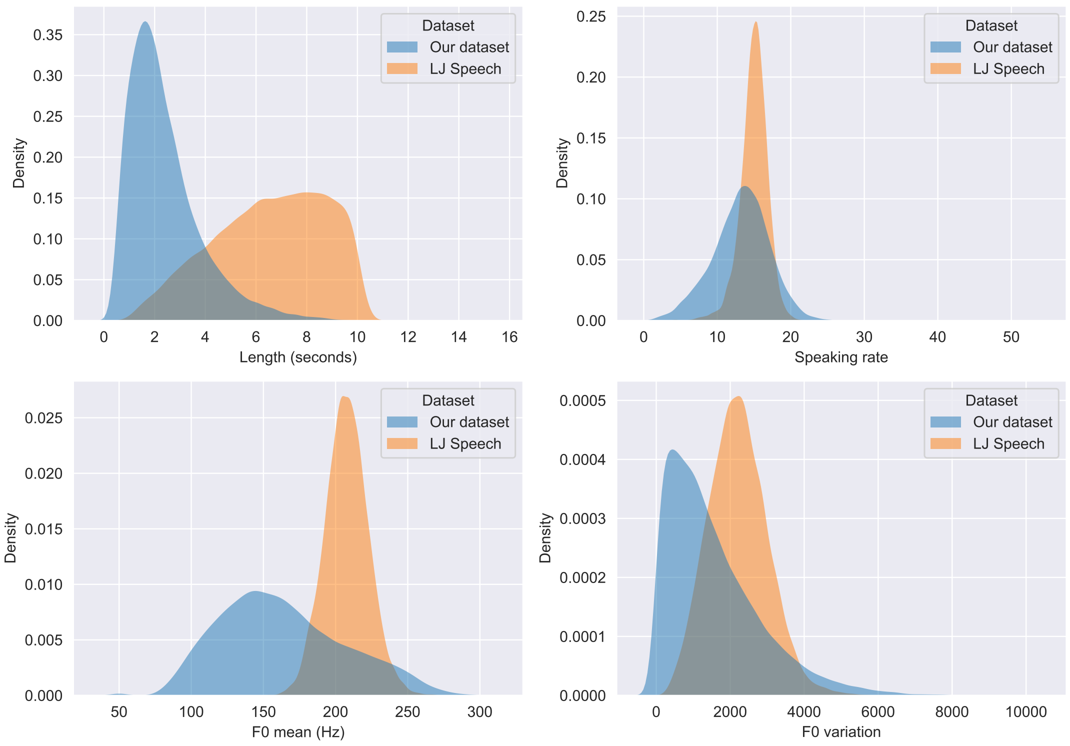


Figure 3.1: Comparison of our dataset and LJ Speech.

3.2.2 Labeling and statistics

The dataset had already been labeled with discrete vocal effort classes for a previous vocal effort classification task. Aside from the split between unvoiced and voiced speech, it could be argued that vocal effort should be considered as a continuous variable. We agree with this assertion but decided to label using discrete classes for the sake of simplicity and interpretability. There is sufficient precedence for discretizing continuous speech attributes in the literature, such as using discrete labels in emotion recognition [76] and quantized articulatory data in speech recognition [77, 78]. We broadly follow the vocal effort labeling precedent set in the literature (see Section 2.1) but add more classes to account for the wide range of vocal effort found in the dataset. It is important to note that while these discrete classes are used for labeling the data, we do not model discrete classes in the VAE-TTS system. We only use these labels to provide feedback regarding how well vocal effort can be disentangled in the latent space. Table 3.1 outlines these classes and their descriptions. We appreciate that these classes are quantitatively ambiguous (except the split between whispered speech and the other classes) and accept that any conclusions reached using these labels need to be considerate of this ambiguity. We hope that as vocal effort becomes a speech attribute

CLASS	NAME	DESCRIPTION
1	Whispered	Unphonated quiet speech with low vocal effort
2	Normal low	Normal phonated speech, lower vocal effort than class 3
3	Normal	Normal speech with ‘average’ vocal effort
4	Lombard	Speech with increased vocal effort but not shouted
5	Shouted	Regular shouted speech
6	Shouted high	Shouting and screaming with the very highest vocal effort

Table 3.1: A description of the vocal effort classes.

explored further in TTS, ASR, and voice conversion, the speech community can begin to formalize labeling methods.

Six different listeners labeled the dataset, and it is worth noting that the labels are relatively noisy. This is perhaps unsurprising; subjectively labeling a continuous variable with categorical classes is prone to disagreement over edge cases. However, this labeling noise is insignificant given the size of the dataset and the fact that we do not use the labels during training. Perhaps more problematically, the distribution of vocal effort classes is highly imbalanced, as shown in Figure 3.2.

To calculate utterance-level energy, duration, fundamental frequency (F_0) mean, and F_0 variation, we used the Parselmouth library [79], a Python wrapper for PRAAT [80]. We also calculated a crude ratio between phonated and unphonated speech by comparing the durations of when the F_0 tracker was and was not able to detect a fundamental frequency (silence removed first). Again, these statistics were not used during training but did help to provide feedback for our experiments, as discussed in Chapters 5 and 6.

3.2.3 Transcription and conversion to phonemes

Approximately three-quarters of the dataset already contained transcriptions. To speed up the process of transcribing the remaining quarter, we first fine-tuned a pre-trained ASR acoustic model (Wav2Vec 2.0 [81]) on our dataset. After inference, any errors in the resulting transcriptions were then manually fixed by hand. These clean transcriptions were then normalized, a process where words such as ‘15’ are converted to ‘fifteen’. We then converted the transcriptions to phonemes using the CMU English Pronouncing Dictionary [82], manually adding out-of-vocabulary words as required

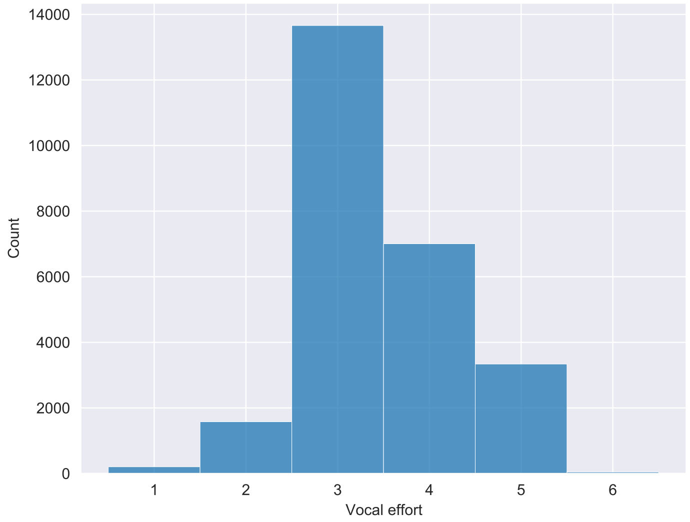


Figure 3.2: Distribution of vocal effort labels.

rather than using a grapheme-to-phoneme model [83, 84]. We had hoped to annotate any non-word vocalizations as part of this process, as this would likely help the sequence-to-sequence model learn alignment between phonemes and acoustics. However, unfortunately, we did not have sufficient time.

Finally, we split the data into training (85%), validation (10%), and testing (5%) sets, ensuring an equal distribution of vocal effort classes across each split.

Chapter 4

Method

4.1 Text-to-speech model

For the core of our system, we use Tacotron 2 [16], a recurrent encoder-attention-decoder model (see Section 2.2 for further background). Tacotron 2 was initially proposed with grapheme input sequences but can also use phoneme sequences, which are often more effective, especially in the English language [85]. These grapheme/phoneme input sequences are fed to a text encoder, and the embedding from this encoder is then passed through three convolutional layers and a bidirectional [86] LSTM [87]. The resulting representation is then passed to a location-sensitive attention network [88] from which a fixed-length context vector is created. This context vector is then consumed by the decoder, a recurrent autoregressive network that predicts mel-spectrograms from these encodings (the Mel scale is a non-linear transform of the spectrogram that emphasizes the lower frequencies, typically more pertinent for speech synthesis). The decoder contains a pre-net through which the prediction from the previous time-step is passed. The output of this pre-net is concatenated with the context vector and passed through two uni-directional LSTMs. The resulting output is concatenated with the context vector and linearly transformed to predict the target mel-spectrogram frame. Finally, a convolutional post-net is used to predict any remaining residual information. Figure 4.1 provides on overview of this architecture.

There are several advantages and disadvantages associated with this model. Tacotron 2 has been shown to produce very natural sounding speech (when coupled with a WaveNet vocoder) and is often held up as a baseline for more recent TTS models. However, the autoregressive nature of Tacotron 2 means that it is slower to train and slower at inference compared to more recent models. More crucially, the attention

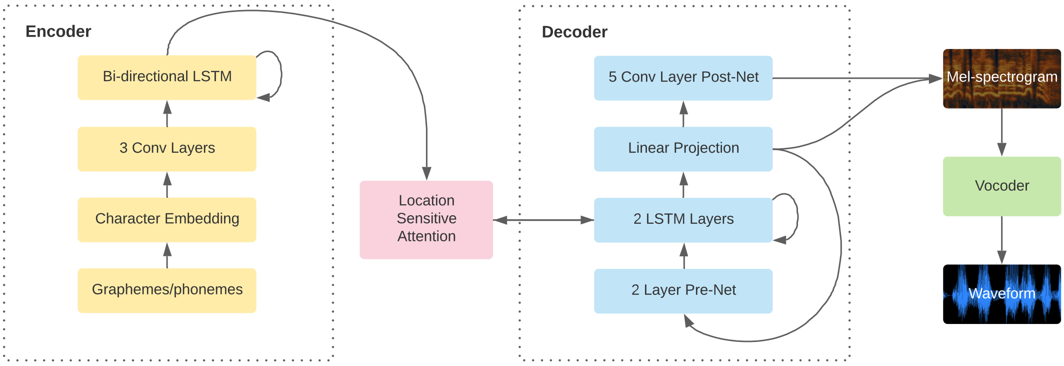


Figure 4.1: Overview of the Tacotron 2 architecture.

mechanism that operates between the text/phonemes and the mel-spectrograms can often be imperfect, leading to skipped or repeated words at inference. More recent approaches replace this attention mechanism with an explicit duration model, leading to improved robustness [19, 69]. Models such as FastSpeech 2 [89] go further and explicitly model pitch, energy, and duration.

However, this explicit modeling is precisely the reason we choose not to use a system such as FastSpeech 2. For our model, we want the VAE to learn latent representations for pitch, energy, and duration, and their relationship to vocal effort, rather than explicitly modeling them individually. In unpublished work by A. Sigurgeirsson et al. (Edinburgh University), the authors find that augmenting a FastSpeech 2 model with a prosody reference encoder [73] was ineffective. The variance predictors appeared to model pitch, energy, and duration so effectively that the reference encoder encoded little useful information. We wish to avoid this and allow the VAE to learn a rich and smoothly regularized representation for vocal effort in an unsupervised manner.

4.2 VAE-Tacotron

To learn the distribution z (see Section 2.3), we first pass the ground-truth mel-spectrogram through a reference encoder. This encoder is the same as the one introduced in [73] and consists of six convolutional layers followed by a Gated Recurrent Unit (GRU) [90]. The outputs of this reference encoder are then passed to two fully connected layers to create the mean and log variance of the distribution z. The dimensionality of z is a hyperparameter that we explore as we seek to discover an appropriate balance between interpretability and disentanglement. We then use the ‘reparameterization trick’ to calculate z from this mean and log variance. This trick is required for backpropogation

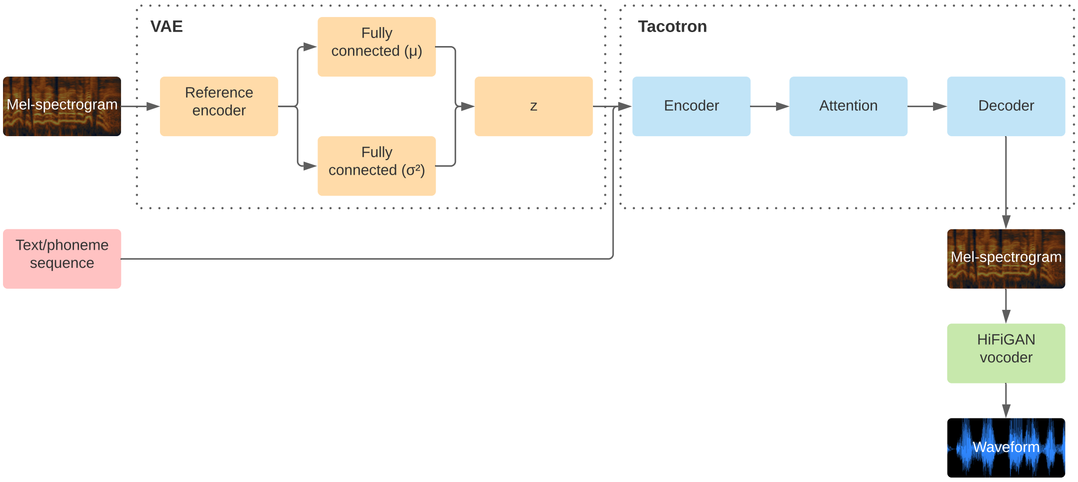
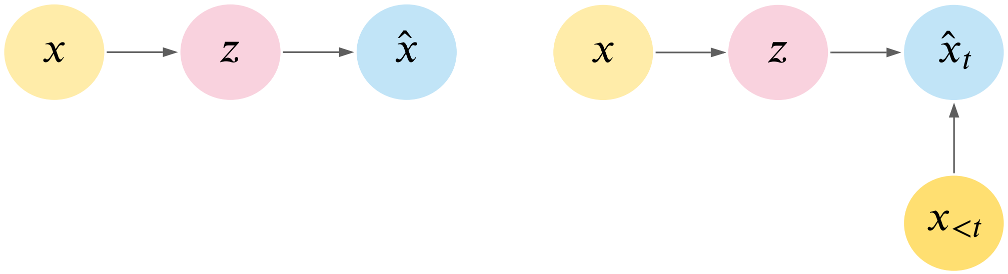


Figure 4.2: Overview of the VAE-Tacotron architecture.

(see [22] for more details). To summarize briefly, to sample z from $\mathcal{N}\left(\mu, \sigma^2\mathbf{I}\right)$ we first sample $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and then calculate $z = \mu + \sigma \odot \varepsilon$ where $\odot$ is the element-wise product. Next, to ensure compatibility with the text/phoneme embedding produced by the Tacotron encoder, we pass z through a fully connected layer with the same dimensionality as this embedding. Then we concatenate these two embeddings together and pass them to the location-sensitive attention network. The context vectors created from this network are then passed to the decoder. Finally, the spectrograms created by the decoded are then vocoded into a waveform using HiFiGAN [91]. We use this generative adversarial network (GAN) [92] vocoder as it is fast at inference and has been demonstrated to yield high-quality audio on a par with autoregressive models such as WaveNet [64]. Figure 4.2 provides an overview of the full VAE-Tacotron architecture.

We do not use the vocal effort labels during training, and the latent distribution z is learned only from observing the mel-spectrograms. We made this decision because we believe there is a risk of failing to learn the true distribution of the data if we rely on discrete (and somewhat noisy) vocal effort labels. However, we accept that there has been some success in conditioning VAE-TTS models in this way [26]. We also do not condition the VAE on the text. We have less defense for this decision, as [27] effectively argues that text does have an affect on the latent posterior. However, for the sake of simplicity, we leave this to be explored in future work.



(a) Typical VAE

(b) VAE plus autoregressive decoder.

Figure 4.3: Decoder pathways.

4.3 KL weight and annealing schedules

With a traditional VAE, the reconstruction $\hat{x}$ is only dependent on z (see Figure 4.3a). By contrast, when we couple a VAE with an autoregressive decoder, $\hat{x}_t$ is dependent on z and $x_{<t}$ (Figure 4.3b). We can consider this as two separate pathways to the decoder, as described in [93]. If the latent representations from z are over-regularized and not rich in information, there is a risk that the decoder will begin to ignore this path during training. This is known as posterior collapse [94] or KL vanishing [93]. In this situation, the variational encoder always produces the same $p(z)$ regardless of the data x, and the decoder always produces $p(\hat{x})$ regardless of z. To alleviate this issue, the most common approach is first to introduce a hyperparameter β (or KL weight) to control the amount of regularization applied to the KL term. Extending Equation 2.3 with this terms gives:

$$\mathcal{L}(x,\hat{x}) + \beta \sum_j KL\left(q_j(z \mid x) \| p(z)\right) \tag{4.1}$$

We start training with the KL weight β at 0, which means the encoder pathway is fully open. This ensures that the latent representations are rich in useful information for the decoder, making it less likely to ignore this path. This configuration is equivalent to a standard autoencoder. To meet the variational objective, we then gradually increase the KL weight over many training steps. By the time the latent space is fully regularized, the representations contained in the latent space are much more likely to be rich in information than if it had been fully regularized from the start, and the decoder is less likely to ignore this path. This annealing approach was introduced in [95] and is a ubiquitous way to alleviate posterior collapse in VAE-decoder systems. However, it is not the only method, and [93] introduces the idea of cyclic annealing of the KL weight. The authors argue that periodically fully re-opening the autoencoder (by setting the KL weight to zero) can lead to even richer latent space representations with improved disentanglement. Figure 5.1 shows a visual comparison of these two annealing schedules. We believe that balancing these two pathways to the decoder is crucial in our experiments and so we explore both of these schedules and a variety of upper KL weight values. However, we do not conduct a deep investigation in this area and appreciate that the techniques for controlling the latent space capacity proposed in [27] would be worth exploring in future work.

4.4 Evaluation of vocal effort control

To evaluate the efficacy of our model, we are primarily concerned with the ability with which we can model vocal effort across a single dimension in the latent space. We focus on this metric for the sake of interpretability and control. Disentangling vocal effort to a single latent should allow us to sample from points across this latent distribution in a simple manner with predictable outcomes. Given that various acoustic attributes such as F_0 mean and variation, energy, and phone duration are correlated with vocal effort [5], we accept that pursuing this goal will result in these ‘lower level’ acoustic variables being entangled. While this is not the typical approach taken with VAE-TTS systems, we believe that for useable and interpretable control of semantically meaningful styles of speech, especially vocal effort, accepting this level of entanglement is appropriate.

Our secondary (and lower priority) evaluation of vocal effort modeling examines how effectively we can transfer vocal effort from a reference utterance. This type of style transfer has been demonstrated with VAE-TTS systems before [24, 28]. While previous experiments have focused on affect, vocal effort transfer should also be possible. We accept that this secondary aim is quite different from our primary aim and that the way in which we optimize for the primary aim may not be optimal for the secondary. However, we still believe that this is a worthwhile line of investigation. If the VAE is capable of modeling vocal effort, even if across a single latent variable, we should expect that it can be also be used for transfer.

To evaluate vocal effort disentanglement and control, we perform various quantitative and subjective evaluations, as detailed in Chapter 5. We also test the efficacy of vocal effort transfer and compare the overall naturalness of our model compared to a baseline.

4.5 Limitations and other considerations

It should be noted that the VAE in this work operates at the utterance level rather than the word, phone, or frame level [29]. We believe that this is a reasonable time domain to model vocal effort, especially for short utterances. However, we do not provide experimental results to back up this claim.

We also concede that this work is focused solely on modeling vocal effort, and no attempt is made to model other important elements of prosody. For example, with this approach, we cannot control whether an utterance has the intonation associated with a question versus a statement. Unlike other work [24, 26, 29], we also do not explore disentangled control of pitch, energy, or speaking rate and only consider these variables in terms of how they relate to vocal effort.

Finally, while we aim to generate high-fidelity natural-sounding speech, we do not focus on this task. We make no attempt to optimize our models for this metric, and for the TTS element of our VAE-TTS system, we rely on the default configuration of the Tacotron 2 model.

Chapter 5

Experiments and Results

5.1 Implementation details

For the following experiments, we modified an open-source VAE-Tacotron model [96], implemented in PyTorch [97]. In each case, we trained the model to 300 epochs (265k steps) using the Adam optimizer [98] with a learning rate of 10^{-3} and weight decay of 10^{-6}. We used a batch size of 24 and a single NVIDIA Quadro RTX 8000 per experiment. Each complete training run took approximately 72 hours, and we ran several experiments simultaneously on separate GPUs. We used an audio sampling rate of 16kHz, which by today's standards may be considered relatively low for TTS. However, we felt that the compromise in audio fidelity was worthwhile for the extra experimentation time made possible by using this lower sampling rate. The remaining Tacotron hyperparameters follow the original paper [16]. Similarly, the hyperparameters for the reference encoder that parameterizes the VAE follow [73].

We initially used grapheme inputs but moved to phoneme inputs after witnessing significant quality and alignment improvements. We also found alignment improvements by pre-training a model on the LJ speech dataset [75]. We used these network weights as a warm-start for our model training. To convert the spectrograms created by the VAE-Tacotron model to audio, we used the HiFiGAN vocoder [91]. We trained this vocoder on the ground-truth audio for 1.57 million steps on a single RTX 8000 GPU (14 days training time).

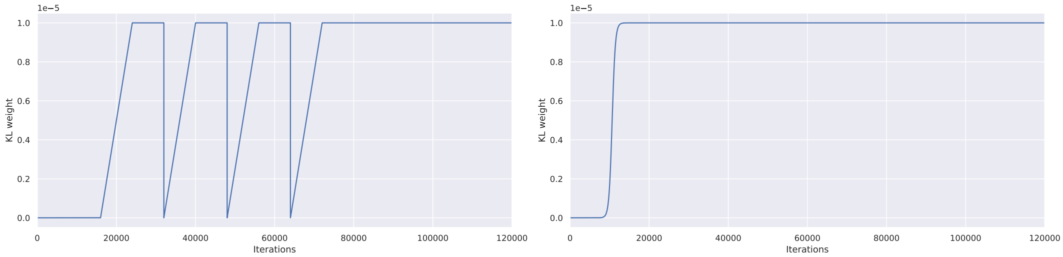


Figure 5.1: Comparison of cyclic and logistic KL annealing schedules

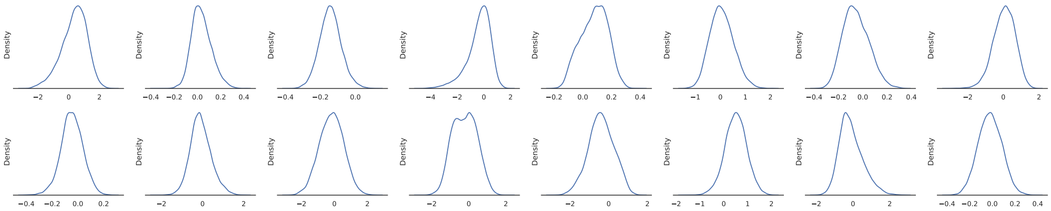


Figure 5.2: Distribution of latent dimensions

5.2 Results

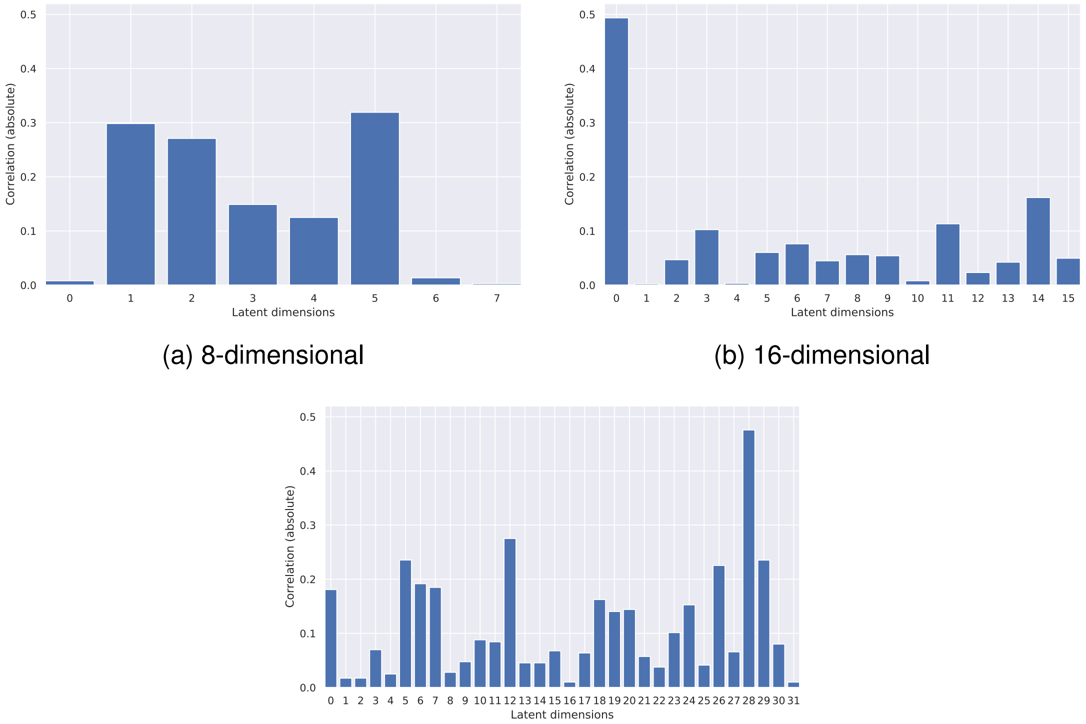
We want to understand how effectively the latent space of the VAE can disentangle vocal effort to a single latent dimension. In order to do so, we pass our dataset through the trained VAE encoder and measure how correlated the means of the latent dimensions are to the vocal effort labels. Unless specified, we only report the single latent dimension with the highest absolute correlation and refer to this simply as ‘vocal effort correlation’.

5.2.1 KL weight and annealing schedule

We experimented with various KL weights and annealing schedules. Figure 5.1 shows the two primary annealing schedules that we used – logistic and cyclic (we also tried a constant KL weight, i.e., no annealing, but found that the model performed poorly in terms of disentanglement). Table 5.1 shows the vocal effort correlation achieved using these annealing schedules and a variety of upper KL weights – that is, the weight at the highest point in the annealing schedule. Using low KL weights can lead to a risk that the latent space of the VAE is not appropriately regulated towards a Gaussian-like distribution. Figure 5.2 shows the distribution of the latent dimensions in our final model. For the remaining experiments, we used the cyclic annealing schedule and an upper KL weight of 10^{-5}.

KL annealing schedule	KL upper weight	Vocal effort correlation
Logistic	10^{-4}	0.21
Cyclic	10^{-4}	0.33
Logistic	10^{-5}	0.46
Cyclic	10^{-5}	**0.49**
Cyclic	10^{-6}	0.45

Table 5.1: Vocal effort correlation with different KL annealing schedules and upper KL weights (16-dimensional VAE).



(a) 8-dimensional

(b) 16-dimensional

(c) 32-dimensional

Figure 5.3: Absolute correlation between vocal effort labels and latent dimensions.

5.2.2 Latent space dimensionality

Figure 5.3 shows the vocal effort correlation across the latent dimensions for three different latent space dimensionalities. Given our desire to control vocal effort with a single latent, we used a 16-dimensional VAE for the remaining experiments. In this case, dimension 0 has an absolute vocal effort correlation of 0.49. Figure 5.4 shows the distribution of the different vocal effort labels across this latent dimension.

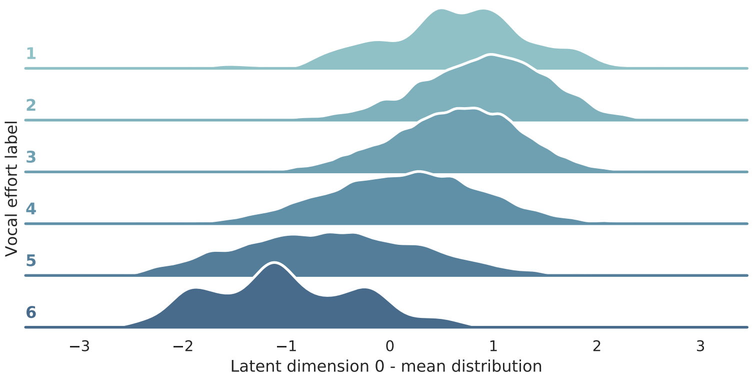
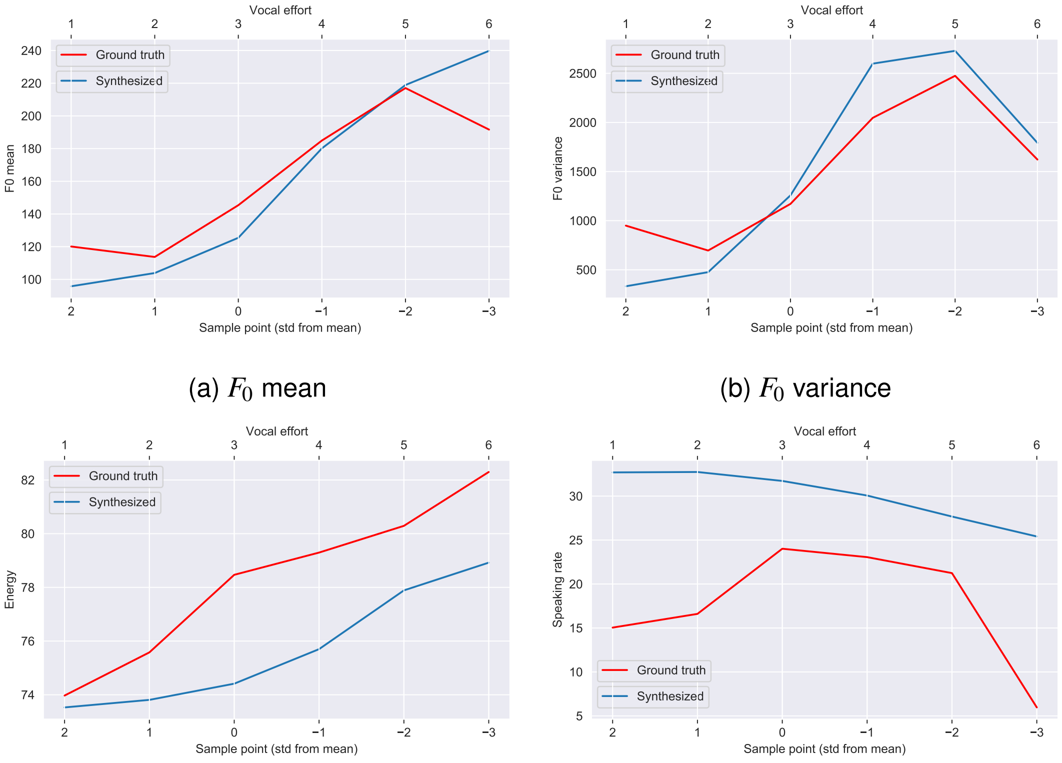


Figure 5.4: Distribution of latent dimension 0 means showing correlation with vocal effort labels.



(a) F_0 mean

(b) F_0 variance

(c) Energy

(d) Speaking rate

Figure 5.5: F_0, energy, and speaking rate correlation to vocal effort labels in the training data and correlation to latent dimension 0 sampling point in the synthesized audio. The upper x-axis is the vocal effort label in the ground truth data and the lower x-axis is the sampling point for the synthesized audio.

LATENT OFFSET	CLASSIFIER PREDICTION	SUBJECTIVE RATING
2σ	2.06 ±0.03	1.92 ±0.05
1σ	2.23 ±0.03	2.44 ±0.06
0	2.57 ±0.04	3.03 ±0.05
-1σ	3.53 ±0.06	3.89 ±0.05
-2σ	4.79 ±0.04	4.72 ±0.05
-3σ	5.14 ±0.03	5.36 ±0.05

Table 5.2: Vocal effort classifier predictions and subjective vocal effort ratings for stimuli generated by sampling across latent dimension 0 (with 95% confidence intervals).

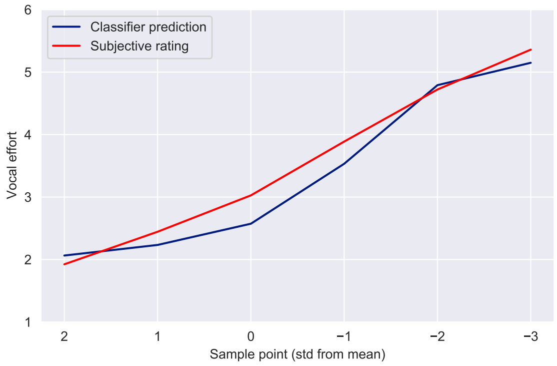
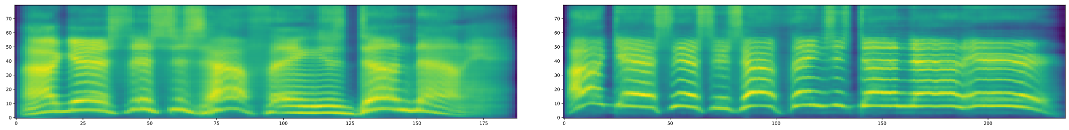


Figure 5.6: Results from Table 5.2 presented visually.

5.2.3 Vocal effort control

With this model, we then explored how effectively dimension 0 could control vocal effort when synthesizing speech. We first set the means of all the VAE dimensions to zero, essentially enforcing a standard multivariate Gaussian across the latent space. We then manipulated latent dimension 0 only, sampling at different standard deviations from its mean. For each sampling point, we generated a sentence taken from the held-out test set. This test set contains 1239 sentences and we sampled from six different points, resulting in 7434 utterances. We then computed utterance-level F_0, energy, and speaking rate statistics from the resulting synthesized audio and compared them to the same statistics computed from the ground-truth training set. Figure 5.5 shows the results.

To provide another quantitative measure, we then ran these generated utterances through a pre-trained vocal effort classifier (see Appendix A for more details). The



(a) Dimension 0 sampled at 0

(b) Dimension 0 sampled at -3σ

Figure 5.7: Two spectrograms generated by sampling from different points along dimension 0. Sentence is: ‘I guess this is how things is done down here’.

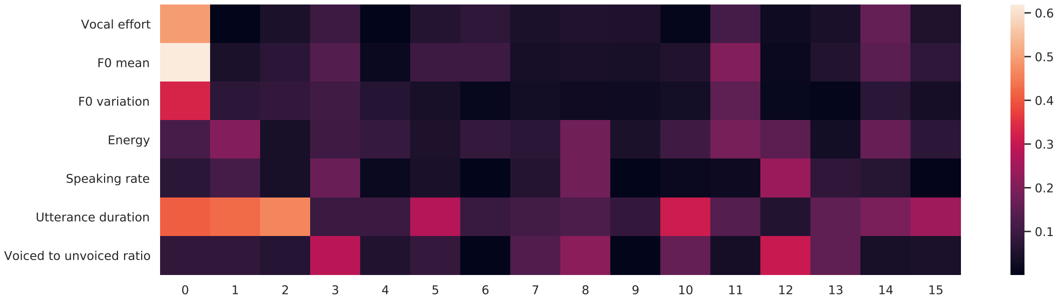


Figure 5.8: Heatmap showing absolute correlation between the latent dimensions and various speech statistics.

results can be seen in Table 5.2.

Finally, we conducted a listening test to subjectively evaluate the vocal effort of the synthesized audio. We first selected twenty sentences from the generated utterances described above. Fifteen evaluators then listened to the six versions of each sentence and were asked to rate each in terms of the perceived vocal effort on a continuous scale from one to six. To help calibrate their choices, six reference utterances from the ground-truth dataset were provided taken from the six vocal effort classes. The evaluators are all audio domain experts and were instructed to wear headphones or use a high-fidelity speaker system. Table 5.2 shows the results from this listener evaluation. Figure 5.7 shows two example spectrograms synthesized by sampling from different points on dimension 0 for one of the test sentences. Audio examples are provided in the supplementary materials.

5.2.4 Naturalness

To ensure that the addition of the VAE element was not having a detrimental effect on the Tacotron model, we trained a baseline model where the VAE element was removed entirely. We then performed listening tests to evaluate the subjective naturalness of the lines. The same fifteen participants were asked to rate a new set of twenty sentences.

MODEL	NATURALNESS
GROUND-TRUTH	4.69 ±0.07
GROUND-TRUTH VOCODED	3.88 ±0.09
VAE-TACOTRON	2.73 ±0.11
BASELINE	1.90 ±0.12

Table 5.3: Mean opinion scores for naturalness (with 95% confidence intervals).

In each case, they were presented with the ground-truth audio at 48kHz, ground-truth at 16kHz passed through the vocoder, a version of the same sentence from the vanilla Tacotron baseline system, and a version from the VAE-Tacotron model where the VAE latent dimensions were set to zero. The participants were asked to give each utterance a naturalness rating from 0 to 5, with 0 being 'not at all human-like and/or noisy' and 5 being 'completely human-like and clear'. The mean opinion scores computed from these ratings can be seen in Table 5.3.

5.2.5 Vocal effort transfer

To test the model's ability to transfer the vocal effort from a reference utterance, we first selected 20 reference utterances per vocal effort class from the test set (120 in total). Next, we passed each of these reference utterances through the VAE encoder to gather its latent representation. These latent representations were then used to condition the synthesis of 20 utterances (with different text content, also taken from the test set). This resulted in 2400 generated utterances, 400 for each vocal effort class. We then passed these utterances through the vocal effort classifier, and the results can be seen in Figure 5.4.

5.2.6 Reproducibility

VAEs are formally non-identifiable, with the result that the latent representations of the VAE may change from one training run to another. To ensure that the model we chose for our evaluations was the not simply the product of random initialization, we re-trained our best performing model with exactly the same hyperparameters. The vocal effort correlation across the latent space can be seen in Figure 5.9 and can be compared to Figure 5.3b.

Reference vocal effort label	Classifier prediction
1	2.66 ±0.11
2	2.22 ±0.05
3	2.44 ±0.08
4	3.38 ±0.11
5	4.27 ±0.11
6	4.11 ±0.13

Table 5.4: Vocal effort classifier predictions for utterances synthesized using the latent representation of a reference utterance (with 95% confidence intervals).

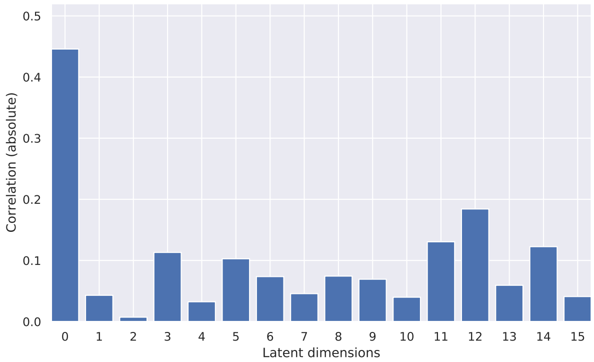


Figure 5.9: Absolute correlation between vocal effort labels and latent dimensions for our evaluation model, re-trained.

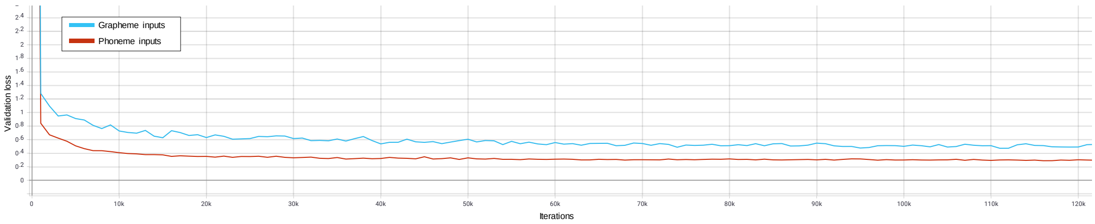


Figure 5.10: Comparison of validation loss for grapheme and phoneme input sequences for the first 120 thousand training iterations.

Chapter 6

Discussion

6.1 TTS with a non-standard dataset

As discussed in Chapter 1, TTS models are typically trained with datasets containing a limited range of stylistic variation, particularly in terms of vocal effort. Therefore, one of our primary objectives was to successfully train a stable model using our dataset, with its wide range of pitch, energy, speaking rate, and vocal effort (see Chapter 3). While there is still room for improvement, this objective has been achieved (we encourage the reader to listen to the provided audio examples).

The most significant contributor to this success would appear to be the addition of the VAE. As discussed in Chapter 1, text-to-speech is a one-to-many mapping problem with many reasonable acoustic expressions of a single sentence of text. This problem is exacerbated when using datasets with a wide range of stylistic variation. Without a suitable mechanism, the vanilla Tacotron model has a single encoder-decoder architecture to model both the lexical and non-lexical information. With a standard loss function, at best, this leads to ‘average’ prosody [21]. With the addition of the VAE, the Tacotron encoder is free to model the explicit input text sequence, leaving the VAE to model the residual non-lexical information [27]. This extra modeling capacity is reflected in the naturalness MOS results comparing the baseline Tacotron model and the VAE-Tacotron model (Table 5.3). We observe that the baseline has considerably worse naturalness. Informally, we found that the baseline model typically produced utterances either with ‘average’ vocal effort (but with poor intelligibility) or with random jumps in vocal effort. The outputs would also typically have far more alignment errors.

We also found that moving from grapheme inputs to phonemes significantly improved the model's ability to learn alignment between these inputs and the audio. The overall validation error also decreased significantly (see Figure 5.10). Although Tacotron can map from graphemes to acoustics, the potential for a single grapheme to have multiple phonetic expressions (particularly in the English language) makes the task harder [85].

Other factors have likely contributed to the successful training of a stable model, but these are harder to quantify with confidence. For example, we initialized the model weights from models pre-trained on a standard TTS dataset (LJ Speech [75]). We found that this sped up convergence and improved alignment, but we provide no quantitative analysis. Similarly, the fact that the dataset we are using is reasonably large (17.2 hours) is almost certainly beneficial. However, again, we did not run experiments with varying amounts of data to investigate this.

Despite the significant improvement of our VAE-TTS model over the baseline, the subjective naturalness is not as high as has been demonstrated with similar systems [24, 26]. We are not certain of the reason for this, but it could be that the highly diverse nature of the dataset is having a negative impact in this regard. We still witness some alignment issues and anticipate improvements by providing richer annotation in the input sequences, particularly regarding non-word vocalizations (breaths, grunts, etc.). The dataset contains many such vocalizations, and appropriately annotating these would likely help the model learn alignment between the input sequences and acoustics. In addition, including annotations for these vocalizations at inference time could yield more naturalistic utterances. Another approach to solving these alignment issues would be to use a different TTS model that does not require an attention mechanism. However, as discussed in Section 4.1, this model would need to be selected carefully to ensure the VAE can still model pitch and duration.

6.2 KL weight and annealing schedule

As discussed in Section 4.3, balancing the KL divergence term is critical to successfully training a VAE in conjunction with a strong autoregressive decoder. We wish to avoid posterior collapse while still regularizing the latent space of the autoencoder towards smooth Gaussian-like distributions. Informally, we found that using a high KL weight did indeed lead to posterior collapse. Table 5.1 shows that KL weights much lower than 1 were required for optimal vocal effort disentanglement, which was somewhat unexpected. However, following the assertions made in [27], we believe that this level of KL term regularization is simply the most appropriate in terms of the information capacity for this task. The risk of a low KL weight is that the latent space is not appropriately regularized to the prior. While the KL divergence in our case is quite high, Figure 5.2 shows that the latent representations are still generally Gaussian-like. It should also be noted that in [28], the authors update their KL divergence loss term every *k* steps in order to avoid posterior collapse. We did not do this, and it may be that much higher KL weights are appropriate when following this scheme.

The differences between the cyclic and logistic KL weight annealing schedules were relatively minor. However, across all the models we trained, cyclic annealing schedules did consistently lead to better vocal effort correlation in the latent space. This supports the arguments made in [93], despite the authors working with text rather than audio. It would appear that fully re-opening the autoencoder pathway (setting the KL weight to zero) several times during training ensured that the decoder did not ignore the information coming from this path and that these ‘restarts’ were beneficial for disentanglement. Overall, while the balancing of the KL term was a crucial element in our experiments, we accept the empirical nature of our results and would seek to explore a more principled approach in the future (see [27]).

6.3 Latent space dimensionality

As seen in Figure 5.3, it is clear that a 16-dimensional latent space provides the clearest disentanglement of vocal effort to a single latent. While the 32-dimensional VAE has strong vocal effort correlation in dimension 28, we can see that vocal effort is also correlated across several other latent dimensions. The use of a 16-dimensional VAE echoes the findings of [24], where the authors claim that a latent space of this dimensionality provided the appropriate balance between disentanglement and interpretability for their particular task. However, to be clear, our findings are almost certainly dependent on numerous variables. These include the size and variation within our dataset, the attribute of speech we seek to isolate and disentangle, and the KL weight and annealing schedule. Unfortunately, at present, we are not aware of a principled approach to selecting the appropriate latent space dimensionality without relying on empirical experiments. It should also be noted that our primary aim was to control vocal effort by sampling across a single latent. For our secondary task of vocal effort transfer, our choice may not be the most optimal as we did not compare latent space dimensionalities for this task. For example, in [28], they find that a 32-dimensional latent space is effective for their task of style transfer.

6.4 Vocal effort control

Our results show that we can effectively model and control vocal effort at inference. As hoped for, we can sample from points across a single latent variable and generate speech with predictable vocal effort. Several different results provide evidence for this claim.

Firstly, we compare various statistics across the ground-truth utterances from the training set and generated utterances from the test set. Figure 5.5 shows the F_0 mean and variance, energy, and speaking rate across these two datasets. For the ground-truth data, we plot these statistics against the vocal effort labels, and for the generated utterances, we plot them against the dimension 0 sampling point. The figures show a strong correlation between these two sets of statistics, particularly with F_0 mean and variance. In terms of energy, we see a strong correlation but at different overall magnitudes. This could be due to the influence of the vocoder, which, depending on its hyperparameters, is likely to produce audio at different amplitudes. Similarly, while the speaking rate between the two datasets follows broadly the same pattern, it is higher overall for the generated utterances. One likely cause is that we do not annotate pauses in our phoneme sequences despite the training set containing such pauses (trimmed to 300ms as discussed in Section 3.2). We also see a lack of speaking rate correlation at the extremes of the vocal effort range. However, the lack of training data at these extremes (44 utterances with vocal effort class 6 and 204 at class 1) make it hard to draw confident conclusions. Informally, we believe this is likely to be because the (very few) vocal effort class 6 utterances in the training set tend to be very short sentences, often single words. When generating high vocal effort utterances from our model, we use the full range of sentence lengths found in the test set.

As well as comparing statistics across these two datasets, we can also reference the comprehensive study on the acoustic effects of vocal effort found in [5]. The authors report correlations between vocal effort and a variety of acoustic variables that are very similar to our findings.

The second quantitative evidence we have for the efficacy of our model is the performance of the vocal effort classifier. As seen in Table 5.2 and Figure 5.6, the classifier monotonically predicts an increase in vocal effort class as we sample across latent dimension 0. To provide assurance of the efficacy of the classifier, Appendix A shows the per-class classification accuracy of this model on a human-labeled test set.

The vocal effort predictions from the classifier are supported by the subjective ratings submitted in the listening test (also Table 5.2). Again, we see a monotonic increase in perceived vocal effort class as we sample across dimension 0. The 95% confidence intervals also remain small despite the relatively small number of listeners and evaluation stimuli. The overall differences between the subjective ratings and the classifier predictions are surprisingly slight, as shown in Figure 5.6.

A key observation is that the classifier and human evaluators rarely, if ever, label an utterance as belonging to class 1 (whispered). This reveals a clear limitation in our work. It would appear that the break between phonated and unphonated speech is not modeled by this one latent dimension. Indeed, looking at Figure 5.8, we can see that the ratio of voiced to unvoiced is correlated quite strongly to latent dimensions other than dimension 0. We are not particularly surprised by this result; as discussed in Section 3.2.2 we believe that while phonated vocal effort should perhaps be considered a continuous variable, the break between phonated and unphonated speech should probably be considered a discrete variable. This limitation may also explain why we could not sample at +3std from the mean on latent dimension 0 while we could sample at -3std from the mean. Theoretically, we would expect that sampling at this point may produce very low vocal effort or even whispered speech; in practice, the model was unable to generate intelligible speech at this sampling point reliably. In future, we hope to model the discrete break in phonated and unphonated speech using a different approach. Potentials routes to explore include the use of a secondary VAE as part of a GMM [24], a learned prior [72], or a hierarchical approach [29].

Another potential reason for this inability to model whispered speech is the lack of training data; our dataset has a highly imbalanced vocal effort distribution (see Figure 3.2). Retraining a model on a smaller but more balanced dataset may reveal the validity of this reasoning.

6.5 Disentanglement

As discussed in Section 2.4, recent research attempting to model expressive speech often focuses on the ability to disentangle various acoustic attributes such as F_0, energy, and phone duration [29]. In our model, we observe that as we sample along dimension 0, F_0 mean and variance, energy, and speaking rate co-correlate and are still very much entangled. However, we argue that if the aim of a model is to creatively control a semantically meaningful ‘higher level’ speech attribute such as vocal effort or affect, we must accept and indeed pursue the appropriate level of entanglement. In natural speech, vocal effort and affect are strongly correlated to a wide range of ‘lower level’, or more granular, acoustic variables [5, 11]. It follows that we would expect the same behavior from our models. Perhaps the most critical consideration when designing latent variable models is to be explicit about the desired attribute to disentangle and then ensure a method is in place to measure success. In our case, while the vocal effort labels were not used during training, they did provide valuable feedback in this regard.

6.6 Vocal effort transfer

Vocal effort transfer was not our primary aim, and this is reflected in our limited evaluation. Referring to Table 5.4, it would appear that the model has some ability to transfer vocal effort from a reference utterance to a synthesized utterance. However, the ambiguity of the results, especially at the lower end of the vocal effort spectrum, suggests that further work is required. Also, we do not compare these vocal effort classifier results to subjective listening test results (our listening test was already a significant length). In informal listening tests, we found that the overall quality (and alignment) was lower than our vocal effort control experiments. This could result from certain latent dimensions corresponding to attributes particular to the reference text sequence that do not match the generated text sequence. (By contrast, in the vocal effort control experiments, we sample from the mean of a normal Gaussian for all dimensions except dimension 0). For example, in Figure 5.8, utterance duration is strongly correlated in dimensions 1 and 2. If there is a significant length mismatch between the reference and generated text, this could be problematic. As mentioned in Section 4.2, [27] suggests that the VAE can be conditioned on the text as well as the audio. Implementing this change may yield improvements for this task. We also anticipate that modifying the model architecture (for example, latent space dimensionality) and optimization (for example, KL weight) for this specific task will likely yield more robust results.

6.7 Naturalness

As seen in Table 5.3, our proposed model generates speech with a reasonable level of naturalness. It was worth observing the result in light of the drop in naturalness from the original ground-truth data to the resampled and vocoded ground-truth data.

This shows that it is likely that the naturalness of our model is compromised simply by the lower acoustic fidelity introduced by the low sample rate and imperfect vocoder. However, we accept that this by no means accounts for the full difference in naturalness between our model and the ground-truth audio. In particular, we found that in informal listening tests, the model particularly struggled with unphonated sections of speech. These less tonal sections tended to sound metallic and 'robotic'. As discussed in Section 4.5, optimizing the model for naturalness and high-fidelity audio was not our primary aim in this work. We expect that simple changes such as increasing the sample rate and training the vocoder for a longer period would improve this result. There are also likely to be a variety of hyperparameter changes we could explore in the Tacotron model, along with the transcript annotation changes discussed earlier in this chapter.

6.8 Other considerations

Our VAE encoding operates at the utterance level, while other approaches attempt to model at the word or phoneme-level [29]. We may find that modeling at these lower time domains yields beneficial results. However, we would argue that the utterance level is the most appropriate for an attribute such as vocal effort, especially when generating short utterances. Also, we would possibly need to incorporate an autoregressive prior across the utterance's VAE encodings to ensure stability and consistency, as proposed in [99].

While we defend the argument in Section 4.2 not to use the vocal effort labels during training, we admit that further investigation is required. The authors of [26] argue that providing supervision (or semi-supervision) during the training of VAE-TTS models can help with reproducibility. Formally, VAEs are non-identifiable [100], and therefore, it can be difficult to predict how the latent representations will form for each training run. Some form of supervision can help to ensure that we can reliably disentangle the relevant speech attributes. However, in our experiments, re-training the evaluation model from scratch resulted in a similar level of vocal effort disentanglement (Figure 5.9). While this does not constitute a thorough investigation, we are satisfied that our model is not simply the product of random initialization.

We also stand by our decision to use a VAE for learning latent representations rather than a simple reference encoder [73] or Global Style Token [74]. Neither of these approaches offers a sampling method that allows smooth interpolation across a latent attribute, and the level of disentanglement is harder to analyze.

Finally, we accept that this work has taken a machine-learning approach to modeling the non-lexical variation in speech. This ‘black-box’ negative definition of prosody is perhaps overly simplistic and ignores much of what we do know about speaking styles. In future, we would seek the valuable contributions of others in the speech research community to help interpret these models.

Chapter 7

Conclusion

In this work, we have demonstrated the ability to model and control vocal effort in a latent variable TTS system. To the best of our knowledge, this is the first time that this speech attribute has been modeled in this manner.

We have shown that it is possible to utilize a dataset containing a significant amount of stylistic and acoustic variation compared to the typical datasets used in research, especially in terms of vocal effort. It would appear that the use of a VAE is highly beneficial in modeling this range of variation and that vanilla TTS systems are unlikely to suffice. Crucially, using a VAE has enabled us to model vocal effort in the latent space, allowing us to easily and effectively generate speech with a wide range of this attribute in an interpretable manner. This interpretability comes at the cost of entanglement of 'lower-level' acoustic variables such as pitch and energy. However, we argue that this trade-off is appropriate for the sake of creative control. As a secondary benefit to our VAE-Tacotron model, we have demonstrated that this approach shows promise in the task of transferring vocal effort from a reference utterance.

7.1 Limitations and future work

Despite these results, there are some significant limitations with the work we have presented. Perhaps the most pertinent is that the dataset we have used is not publicly available. Given the lack of similar datasets, this will make it hard to reproduce our results. To this end, one key direction of future work is to develop a publicly available speech dataset with a wide range of vocal effort. Given the costs of developing such a dataset, it is unlikely to be as large as the one used in this work. However, recording a more balanced vocal effort dataset may mean that a smaller size will suffice. Coupled with insights from the field of low-resource TTS [101–103], we are hopeful that it should be possible to replicate and build on our results.

Another clear limitation to our work is the inability to model whispered speech. In the future, we would wish to explore different approaches to modeling this discrete split in phonation. Promising directions to explore include the use of GMMs [24], hierarchical VAEs [29], and learned priors [72]. As part of this work, we would also seek to explore multi-speaker modeling and a more thorough investigation into the most appropriate time domain for the VAE. Our model operates at the utterance level, but the phoneme or frame level may be more appropriate.

Furthermore, our work also makes no attempt to model or explain any feature of prosody beyond vocal effort. While this attribute is an important source of variation in speech, a robust TTS system would require interpretable control over a range of other prosodic variables.

Our model also still suffers from alignment issues. As discussed in Section 6.1, annotating non-word vocalizations is one possible route to improvement. Another solution is to experiment with systems that do not rely on attention mechanisms and instead explicitly model duration. However, vocal effort is highly correlated to variables such as pitch, energy, and duration. As discussed in Section 4.1, explicitly modeling these attributes out-with the VAE will make it harder, if not impossible, for the latent space of the VAE to encode meaningful vocal effort representations. Thus, we believe that designing robust latent variable TTS systems is still an open research question.

Our vocal effort transfer results are promising but not conclusive. However, given that [28] has demonstrated the ability to transfer prosodic attributes using VAE-TTS models, we expect that we could achieve a similar level of robustness for vocal effort with architectural and optimization modifications.

Although we have attempted to provide a light evaluation of reproducibility across multiple training runs, a lack of predictability is still inherent in training VAEs. While we have argued the case for not using labels during training, providing some form of light supervision as proposed in [26] may alleviate this issue. Likewise, we would wish to investigate a more principled approach to the choice of latent space dimensionality, KL weight, and KL annealing schedule [27].

Despite these limitations, we believe this work represents a valuable contribution to the field of expressive speech synthesis. We hope that exploring vocal effort in this context, along with other domains such as ASR and voice conversion, will continue to grow into an active area of research.

Bibliography

[1] Albert Mehrabian et al. *Silent messages*. Vol. 8. 152. Wadsworth Belmont, CA, 1971.

[2] Paul Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 19, 2009. 626 pp. ISBN: 978-0-521-89927-7.

[3] D Robert Ladd. *Intonational phonology*. Cambridge University Press, 2008.

[4] Michael Wagner and Duane G. Watson. "Experimental and theoretical advances in prosody: A review". *Language and cognitive processes* 25.7 (1, 2010), pp. 905–945. ISSN: 0169-0965.

[5] Hartmut Traunmüller and Anders Eriksson. "Acoustic effects of variation in vocal effort by men, women, and children". *The Journal of the Acoustical Society of America* 107.6 (2000), pp. 3438–3451. ISSN: 0001-4966.

[6] Anders Eriksson and Hartmut Traunmüller. "Perception of vocal effort and distance from the speaker on the basis of vowel utterances". *Perception & psychophysics* 64.1 (2002), pp. 131–139.

[7] Harlan Lane and Bernard Tranel. "The Lombard sign and the role of hearing in speech". *Journal of Speech and Hearing Research* 14.4 (1971), pp. 677–709.

[8] Jean-Claude Junqua. "The Lombard reflex and its role on human listeners and automatic speech recognizers". *The Journal of the Acoustical Society of America* 93.1 (1993), pp. 510–524.

[9] John W Black. "The Effect of Room Characteristics upon Vocal Intensity and Rate". *The Journal of the Acoustical Society of America* 21.4 (1949), pp. 461–461.

[10] David Pelegrin-Garcia et al. "Vocal effort with changing talker-to-listener distance in different acoustic environments". *The Journal of the Acoustical Society of America* 129.4 (2011), pp. 1981–1990.

[11] Klaus R Scherer. “Vocal affect expression: a review and a model for future research.” *Psychological bulletin* 99.2 (1986), p. 143.

[12] Christer Gobl and Ailbhe Nı Chasaide. “The role of voice quality in communicating emotion, mood and attitude”. *Speech communication* 40.1-2 (2003), pp. 189–212.

[13] Kei Fujii, Hideki Kashioka, and Nick Campbell. “Target cost of f0 based on polynomial regression in concatenative speech synthesis”. *Proc. ICPhS*. 2003, pp. 2577–2580.

[14] Takayoshi Yoshimura et al. “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis”. *Sixth European Conference on Speech Communication and Technology*. 1999.

[15] Yuxuan Wang et al. “Tacotron: Towards End-to-End Speech Synthesis”. *Proc. Interspeech 2017* (2017), pp. 4006–4010.

[16] Jonathan Shen et al. “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions”. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 4779–4783.

[17] Wei Ping et al. “Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning”. *International Conference on Learning Representations*. 2018.

[18] Yaniv Taigman et al. “VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop”. *International Conference on Learning Representations*. 2018.

[19] Yi Ren et al. “FastSpeech: fast, robust and controllable text to speech”. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2019, pp. 3171–3180.

[20] Naihan Li et al. “Neural speech synthesis with transformer network”. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 6706–6713.

[21] Zack Hodari, Oliver Watts, and Simon King. “Using generative modelling to produce varied intonation for speech synthesis”. *Proc. 10th ISCA Speech Synthesis Workshop*, pp. 239–244.

[22] Diederik P. Kingma and Max Welling. “Auto-Encoding Variational Bayes”. *arXiv:1312.6114 [cs, stat]* (2014).

[23] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. “Stochastic Backpropagation and Approximate Inference in Deep Generative Models”. *International Conference on Machine Learning*. PMLR, 2014, pp. 1278–1286.

[24] Wei-Ning Hsu et al. “Hierarchical Generative Modeling for Controllable Speech Synthesis”. *International Conference on Learning Representations*. 2018.

[25] Gustav Eje Henter et al. “Deep Encoder-Decoder Models for Unsupervised Learning of Controllable Speech Synthesis”. *arXiv:1807.11470 [cs, eess, stat]* (2018).

[26] Raza Habib et al. “Semi-Supervised Generative Modeling for Controllable Speech Synthesis”. *International Conference on Learning Representations*. 2019.

[27] Eric Battenberg et al. “Effective Use of Variational Embedding Capacity in Expressive End-to-End Speech Synthesis”. *arXiv:1906.03402 [cs, eess]* (2019).

[28] Ya-Jie Zhang et al. “Learning latent representations for style control and transfer in end-to-end speech synthesis”. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 6945–6949.

[29] Guangzhi Sun et al. “Fully-Hierarchical Fine-Grained Prosody Modeling For Interpretable Speech Synthesis”. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 6264–6268.

[30] Heiga Zen et al. “LibriTTS: A corpus derived from LibriSpeech for text-to-speech”. *Proc. Interspeech 2019* (2019), pp. 1526–1530.

[31] Simon King and Vasilis Karaiskos. “The Blizzard Challenge 2013”. *Blizzard Challenge Workshop*. 2013.

[32] John Kominek and Alan W Black. “The CMU Arctic speech databases”. *Fifth ISCA workshop on speech synthesis*. 2004.

[33] Christophe Veaux, Junichi Yamagishi, and Simon King. “The voice bank corpus: Design, collection and data analysis of a large regional accent speech database”. *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation*. IEEE. 2013, pp. 1–4.

[34] Ann Clifton et al. “The Spotify Podcast Dataset”. *arXiv:2004.04270 [cs]* (2020).

[35] Jort F. Gemmeke et al. "Audio Set: An ontology and human-labeled dataset for audio events". *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 776–780.

[36] Anthony Rousseau, Paul Deléglise, and Yannick Esteve. "TED-LIUM: an Automatic Speech Recognition dedicated corpus." *LREC*. 2012, pp. 125–129.

[37] Soujanya Poria et al. "Meld: A multimodal multi-party dataset for emotion recognition in conversations". *arXiv preprint arXiv:1810.02508* (2018).

[38] Marc Schröder and Martine Grice. "Expressing Vocal Effort in Concatenative Synthesis". *Proceedings of the 15th International Conference of Phonetic Sciences* (2003).

[39] Oytun Turk et al. "Voice quality interpolation for emotional text-to-speech synthesis". *Ninth European Conference on Speech Communication and Technology*. 2005.

[40] Tuomo Raitio et al. "Analysis of HMM-based Lombard speech synthesis". *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[41] Tuomo Raitio et al. "Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise". *Computer Speech & Language* 28.2 (2014), pp. 648–664. ISSN: 0885-2308.

[42] Tuomo Raitio et al. "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort". *Fifteenth Annual Conference of the International Speech Communication Association*. 2014.

[43] Bajibabu Bollepalli, Lauri Juvela, Paavo Alku, et al. "Lombard Speech Synthesis Using Transfer Learning in a Tacotron Text-to-Speech System." *Interspeech*. 2019, pp. 2833–2837.

[44] Qiong Hu et al. "Whispered and Lombard Neural Speech Synthesis". *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2021, pp. 454–461.

[45] T. Raitio et al. "Analysis and synthesis of shouted speech". *INTERSPEECH*. 2013.

[46] E. Holmberg, R. Hillman, and J. Perkell. "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice." *The Journal of the Acoustical Society of America* (1988).

[47] Ingo R Titze and Daniel W Martin. *Principles of voice production*. Acoustical Society of America, 1998.

[48] Eric J Hunter et al. "Toward a consensus description of vocal effort, vocal load, vocal loading, and vocal fatigue". *Journal of Speech, Language, and Hearing Research* 63.2 (2020), pp. 509–532.

[49] Ana Ramirez López et al. "Speaking Style Conversion from Normal to Lombard Speech Using a Glottal Vocoder and Bayesian GMMs." *Interspeech*. 2017, pp. 1363–1367.

[50] Shreyas Seshadri et al. "Cycle-consistent Adversarial Networks for Non-parallel Vocal Effort Based Speaking Style Conversion". *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, 2019, pp. 6835–6839. ISBN: 978-1-4799-8131-1.

[51] M. Cotescu et al. "Voice Conversion for Whispered Speech Synthesis". *IEEE Signal Processing Letters* 27 (2020), pp. 186–190. ISSN: 1558-2361.

[52] Chi Zhang and John HL Hansen. "Analysis and classification of speech mode: whispered through shouted". *Eighth Annual Conference of the International Speech Communication Association*. 2007.

[53] Santi Prieto et al. "Shouted Speech Compensation for Speaker Verification Robust to Vocal Effort Conditions". *Interspeech 2020*. Interspeech 2020. ISCA, 25, 2020, pp. 1511–1515.

[54] Cecil H Coker. "A model of articulatory dynamics and control". *Proceedings of the IEEE* 64.4 (1976), pp. 452–460.

[55] Christine H Shadle and Robert I Damper. "Prospects for articulatory synthesis: A position paper". *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*. 2001.

[56] P Seeviour, J Holmes, and M Judd. "Automatic generation of control signals for a parallel formant speech synthesizer". *ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE. 1976, pp. 690–693.

[57] Dennis H Klatt. "Software for a cascade/parallel formant synthesizer". *the Journal of the Acoustical Society of America* 67.3 (1980), pp. 971–995.

[58] Eric Moulines and Francis Charpentier. “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”. *Speech communication* 9.5-6 (1990), pp. 453–467.

[59] Andrew J Hunt and Alan W Black. “Unit selection in a concatenative speech synthesis system using a large speech database”. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Vol. 1. IEEE. 1996, pp. 373–376.

[60] Keiichi Tokuda et al. “Speech parameter generation algorithms for HMM-based speech synthesis”. *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*. Vol. 3. IEEE. 2000, pp. 1315–1318.

[61] Heiga Zen, Keiichi Tokuda, and Alan W Black. “Statistical parametric speech synthesis”. *speech communication* 51.11 (2009), pp. 1039–1064.

[62] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. *arXiv preprint arXiv:1409.0473* (2014).

[63] Ashish Vaswani et al. “Attention is all you need”. *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[64] Aaron van den Oord et al. “Wavenet: A generative model for raw audio”. *arXiv preprint arXiv:1609.03499* (2016).

[65] Kundan Kumar et al. “Melgan: Generative adversarial networks for conditional waveform synthesis”. *arXiv preprint arXiv:1910.06711* (2019).

[66] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. “Waveglow: A flow-based generative network for speech synthesis”. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 3617–3621.

[67] Nanxin Chen et al. “WaveGrad: Estimating gradients for waveform generation”. *arXiv preprint arXiv:2009.00713* (2020).

[68] Tom Kenter et al. “CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network”. *International Conference on Machine Learning*. PMLR. 2019, pp. 3331–3340.

[69] Adrian Łańcucki. "Fastpitch: Parallel text-to-speech with pitch prediction". *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 6588–6592.

[70] Irina Higgins et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". *International Conference on Learning Representations*. 2017.

[71] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. "Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder". *Proc. Interspeech 2018* (2018), pp. 3067–3071.

[72] Penny Karanasou et al. "A learned conditional prior for the VAE acoustic space of a TTS system". *arXiv:2106.10229 [cs, eess]* (14, 2021). arXiv: `2106.10229`.

[73] RJ Skerry-Ryan et al. "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron". *International Conference on Machine Learning*. PMLR. 2018, pp. 4693–4702.

[74] Yuxuan Wang et al. "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis". *International Conference on Machine Learning*. PMLR. 2018, pp. 5180–5189.

[75] Keith Ito and Linda Johnson. *The LJ Speech Dataset*. `https://keithito.com/LJ-Speech-Dataset/`. 2017.

[76] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases". *Pattern Recognition* 44.3 (2011), pp. 572–587. ISSN: 0031-3203.

[77] Simon King et al. "Speech production knowledge in automatic speech recognition". *The Journal of the Acoustical Society of America* 121.2 (2007), pp. 723–742. ISSN: 0001-4966.

[78] T.A. Stephenson, M.M. Doss, and H. Bourlard. "Speech recognition with auxiliary information". *IEEE Transactions on Speech and Audio Processing* 12.3 (2004), pp. 189–203. ISSN: 1558-2353.

[79] Yannick Jadoul, Bill Thompson, and Bart de Boer. "Introducing Parselmouth: A Python interface to Praat". *Journal of Phonetics* 71 (2018), pp. 1–15.

[80] Paul Boersma and David Weenink. *Praat: doing phonetics by computer [Computer program].* Version 6.1.38, retrieved 2 January 2021 `http://www.praat.org/`. 2021.

[81] Alexei Baevski et al. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". *Advances in Neural Information Processing Systems* 33 (2020).

[82] Carnegie Mellon University. *The CMU Pronouncing Dictionary (version 0.7b).* `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`. 2014.

[83] Maximilian Bisani and Hermann Ney. "Joint-sequence models for grapheme-to-phoneme conversion". *Speech communication* 50.5 (2008), pp. 434–451.

[84] Kanishka Rao et al. "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks". *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE. 2015, pp. 4225–4229.

[85] Jason Fong et al. "A Comparison of Letters and Phones as Input to Sequence-to-Sequence Models for Speech Synthesis". *10th ISCA Speech Synthesis Workshop.* 10th ISCA Speech Synthesis Workshop. ISCA, 20, 2019, pp. 223–227.

[86] Mathias Berglund et al. "Bidirectional recurrent neural networks as generative models". *Advances in Neural Information Processing Systems* 28 (2015), pp. 856–864.

[87] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". *Neural computation* 9.8 (1997), pp. 1735–1780.

[88] Jan K Chorowski et al. "Attention-Based Models for Speech Recognition". *Advances in Neural Information Processing Systems.* Vol. 28. Curran Associates, Inc., 2015.

[89] Yi Ren et al. "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech". International Conference on Learning Representations. 28, 2020.

[90] Junyoung Chung et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". *arXiv preprint arXiv:1412.3555* (2014).

[91] 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* 33 (2020).

[92] Ian Goodfellow et al. “Generative adversarial nets”. *Advances in neural information processing systems* 27 (2014).

[93] Hao Fu et al. “Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing”. *NAACL-HLT (1)*. 2019.

[94] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. “Neural Discrete Representation Learning”. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 6309–6318.

[95] Samuel R Bowman et al. “Generating sentences from a continuous space”. *20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016*. Association for Computational Linguistics (ACL). 2016, pp. 10–21.

[96] Jinhan Choi. *tacotron2-vae*. `https://github.com/jinhan/tacotron2-vae`. 2019.

[97] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library”. *Advances in neural information processing systems* 32 (2019), pp. 8026–8037.

[98] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. *arXiv:1412.6980 [cs]* (2017).

[99] Guangzhi Sun et al. “Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior”. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 6699–6703.

[100] Aapo Hyvärinen and Petteri Pajunen. “Nonlinear independent component analysis: Existence and uniqueness results”. *Neural networks* 12.3 (1999), pp. 429–439.

[101] Yuan-Jui Chen et al. “End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning”. *Proc. Interspeech 2019* (2019), pp. 2075–2079.

[102] Yu-An Chung et al. “Semi-supervised training for improving data efficiency in end-to-end speech synthesis”. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 6940–6944.

[103] Jin Xu et al. "Lrspeech: Extremely low-resource speech synthesis and recognition". *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 2802–2812.

[104] Kaiming He et al. "Deep residual learning for image recognition". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[105] Daniel S Park et al. "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition". *Proc. Interspeech 2019* (2019), pp. 2613–2617.

[106] Joel Shor et al. "Towards learning a universal non-semantic representation of speech". *Proc. Interspeech 2020* (2020), pp. 140–144.

Appendix A

Vocal effort classifier

For a previous vocal effort classification task, we trained a vocal effort classifier on a corpus of over 30 thousand labeled utterances. Initially, we used a ResNet architecture [104] operating on the spectrograms coupled with the data augmentation method SpecAugment [105]. However, we witnessed significant improvements by instead using embeddings extracted from a pre-trained self-supervised model, TRILL [106]. These embeddings were then passed through a simple linear classifier. The per-class results for a held-out test set can be seen in Table A.1.

VOCAL EFFORT CLASS	ACCURACY (%)
1	79.7
2	89.5
3	84.3
4	67.1
5	83.2
6	96.2
OVERALL	81.6

Table A.1: Per-class vocal effort classifier accuracy on a held-out test set.