SEMI-SUPERVISED GENERATIVE MODELING FOR CONTROLLABLE SPEECH SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a novel generative model that combines state-of-the-art neural text-to-speech (TTS) with semi-supervised probabilistic latent variable models. By providing partial supervision to some of the latent variables, we are able to force them to take on consistent and interpretable purposes, which previously hasn't been possible with purely unsupervised TTS models. We demonstrate that our model is able to reliably discover and control important but rarely labelled attributes of speech, such as affect and speaking rate, with as little as 1% (30 minutes) supervision. Even at such low supervision levels we do not observe a degradation of synthesis quality compared to a state-of-the-art baseline. We will release audio samples at https://tts-demos.github.io/.

1 INTRODUCTION

The ability to reliably control high level attributes of of speech, such as emotional expression (affect) or speaking rate, is often desirable in speech synthesis applications. Achieving this control however is made difficult by the necessity of acquiring a large quantity of high quality labels. In this paper we show that semi-supervised latent variable models can take us a significant step closer towards solving this problem.

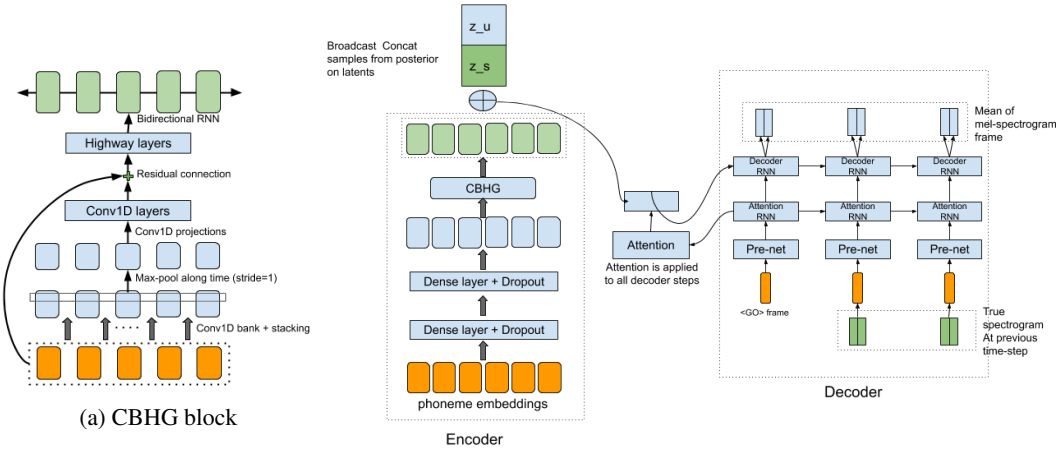
Combining state-of-the-art neural text-to-speech (TTS) systems with probabilistic latent variable models provides a natural framework for discovering aspects of speech that are rarely labelled or even difficult to describe. Both inferring the latent prosody and generating samples with sufficient variety requires reasoning about uncertainty and is thus a natural fit for deep generative models.

There has been recent progress in applying stochastic gradient variational Bayes (SGVB) (Kingma & Welling, 2013; Rezende et al., 2014) to training probabilistic neural TTS models. Battenberg et al. (2019) and Hsu et al. (2018) have shown that it is possible to use latent variable models to discover features such as speaking style, speaking rate, arousal, gender and even the quality of the recording environment.

However, these models are formally non-identifiable (Hyvärinen & Pajunen, 1999) and this implies that repeated training runs will not reliably discover the same latent attributes. Even if they did, a lengthy human post-processing stage is necessary to identify what the model has learned on any given training run. In order to be of practical use for control, it is not enough for the models to discover latent attributes, they need to do so reliably and in a way that is robust to random initialization and to changes in the model. We demonstrate that the addition of even modest amounts of supervision can be sufficient to achieve this reliability.

By augmenting state-of-the art neural TTS with semi-supervised deep generative models within the VAE framework (Kingma et al., 2014; Narayanaswamy et al., 2017), we show that it is possible to not only discover latent attributes of speech but to do so in a reliable and controllable manner. In particular we are able to achieve reliable control over affect, speaking rate and F0 variation (F0 is the fundamental frequency of oscillation of the vocal folds). Further, we provide demonstrations that it is possible to transfer controllability to speakers for whom we have no labels. Our core contributions are:

- To combine semi-supervised latent variable models with Neural TTS systems, producing a system that can *reliably* discover attributes of speech we wish to control.



(a) CBHG block

(b) seq-to-seq network.

Figure 1: Schematic showing how we parameterize the conditional likelihood $p(x|y, z_u, z_s)$. Left: A block of 1-d convolutions and RNNs originally introduced by Wang et al. (2017). Right: Schematic of the sequence-to-sequence network that outputs the means of our auto-regressive distribution. At each decoder time step, the network outputs the means for the next two spectrogram frames.

- To demonstrate that as little as 1% supervision can be sufficient to improve prosody and allow control over speaking rate, fundamental frequency (F0) variation and affect, a problem of interest to the speech community for well over two decades (Schröder, 2001).
- To embue TTS models with control over affect, F0 and speaking rate whilst still maintaining prosodic variation when sampling.

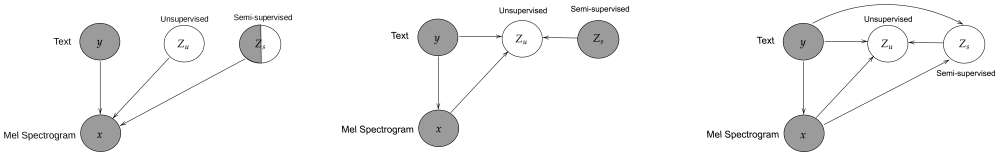
2 Generative Model

Our generative model, shown in figures 1 and 2a, consists of an autoregressive distribution over a sequence of acoustic features, $x_{1...t}$, that are generated conditioned on a sequence of text, $y_{1...k}$, and on two latent variables, z_u and z_s. The latent variables can be discrete or continuous. z_s represents the variations in prosody that we seek to control and is semi-supervised. z_u is fully unobserved and represents latent variations in prosody (intonation, rhythm, stress) that we wish to model but not explicitly control. Once trained, our model can be used to synthesize acoustic features from text. Similar to Tacotron 2 (Shen et al., 2018), we then generate waveforms by training a second network such as WaveNet (van den Oord et al., 2016) or WaveRNN (Kalchbrenner et al., 2018) to act as a vocoder.

We parameterize our likelihood $p(x_{1...t}|y_{1...t}, z_u, z_s, \theta)$ by a sequence-to-sequence neural network with attention (Shen et al., 2018; Graves, 2013; Bahdanau et al., 2014) that is shown schematically in figure 1. Details largely follow Tacotron (Wang et al., 2017) and are given in appendix A. At each time step we model a mel-spectrogram frame with a fixed variance isotropic Laplace distribution whose mean is output by the neural network. We condition each of the latent variables by concatenating the vectors z_u and z_s to the representation of the text-encoder, before the application of the attention mechanism. In the case of continuous z we use a standard normal prior and in the case of discrete z we use a uniform categorical prior with one-hot encoding.

2.1 Semi-Supervised Training

Following Kingma et al. (2014); Narayanaswamy et al. (2017), we train our model via stochastic gradient variational Bayes (SGVB). That is we approximately maximize the log-likelihood of our training data by maximizing a variational lower bound using stochastic gradient ascent. Since we are training with semi-supervision we in fact need two lower bounds: one for the data points for which z_s is observed; one for the case where z_s is unobserved. In our models the fully latent variable z_u is



(a) Generative model (b) Supervised posterior (c) Unsupervised posterior

Figure 2: Left: The graphical model showing the conditional independence assumptions between each of the stochastic variables. Centre: The structure of the variational distribution used to approximate the posterior for supervised data points and Right: fully unsupervised points.

always continuous but the semi-supervised latent z_s can be continuous or discrete. The conditional independence structure of our variational distributions is shown in figures 2b and 2c. On supervised data, the per-datapoint bound is:

$$\begin{aligned}
\log p(x, z_s|y) &= \log \int p(x, z_u, z_s|y, \theta) dz_u \\
&\geq E_{q(z_u|x,y,z_s,\phi)} \left[\log \left(\frac{p(x|y, z_u, z_s, \theta) p(z_u) p(z_s)}{q(z_u|x, y, z_s, \phi)} \right) \right] \\
&= E_{q(z_u|x,y,z_s,\phi)} \left[\log p(x|y, z_u, z_s, \theta) \right] + \log p(z_s) - D_{KL}(q(z_u|x, y, z_s, \phi) \| p(z_u)) \\
&= \mathcal{L}_s(\theta, \phi)
\end{aligned}$$

Where $q(z_u|x, y, z_s, \phi)$ is a parametric variational distribution introduced to approximately marginalize z_u. θ are the parameters of the generative model and ϕ are the parameters of the variational distributions. The intractable integrals are approximated with reparameterized samples. For the cases where z_s is unobserved and discrete, the bound is:

$$\log p(x|y) = \log \int \sum_{z_s} p(x, z_u, z_s|y) dz_u \tag{1}$$

$$\geq \sum_{z_s} \left[q(z_s|x, y, \phi) \mathcal{L}_s(\theta, \phi) \right] + H(q(z_s|x, y, \phi)) \tag{2}$$

$$= \mathcal{L}_u(\theta, \phi) \tag{3}$$

and when z_s is continuous we replace the sum above with an integral and again approximate with reparameterized samples. The variational distributions are parameterized by a neural network that takes as input the text, spectrograms and other conditioning variables and outputs the parameters of the distribution. The exact structure of this network is given in appendix A. We have implicitly assumed that $q(z_u, z_s|x, y, \phi)$ may be factorized as $q(z_u, z_s|x, y, \phi) = q(z_u|x, y, z_s, \phi) q(z_s|x, y, \phi)$ with shared parameters between these two distributions (see appendix A). Optimizing the variational objective with respect to the parameters ϕ encourages the variational distributions to match the posterior of the generative model $p(z_u, z_s|x, y, \theta)$. Unlike previous work (Hsu et al., 2018), we do not assume that the posterior on the latents is independent of the text, as this dependence likely exists in the model due to explaining away. That is to say that although the text and the latents are independent in our prior, observing the spectrogram correlates them in the posterior because they both explain variation in the spectrogram. This has been shown to be significant by (Battenberg et al., 2019).

If we define:

$$\tilde{q}(z_s|x, y) = \begin{cases} q(z_s|x, y, \phi) & \text{if unsupervised} \\ \gamma \delta(z_s - z_{s_{observed}}) & \text{if supervised} \end{cases} \tag{4}$$

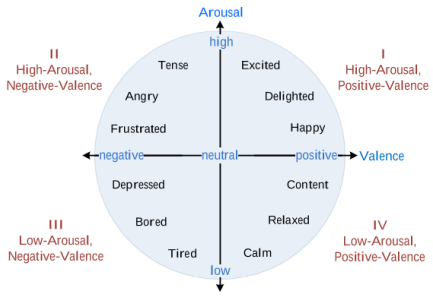


Figure 3: The circumplex model of emotion. Each possible emotion is represented in a 2 dimensional plane consisting of an arousal dimension and valence dimension. This figure is borrowed from Munoz-de Escalona & Canas (2017).

then we can write the overall objective over both the supervised and unsupervised points succinctly as [1]:

$$\mathcal{L}(\theta, \phi) = E_{x,y,z_s}\left[\sum_{z_s}\left[\tilde{q}(z_s|x, y, \phi)\mathcal{L}_s(\theta, \phi)\right] + H(\tilde{q}(z_s|x, y, \phi))\right] \tag{5}$$

Where summation would again be replaced by integration for continuous z_s and γ (shown in equation 4) is a weighting factor that pre-multiplies the loss for any supervised point. This weighting was also used in previous work such as Narayanaswamy et al. (2017), who showed it to be beneficial at very low levels of supervision.

Writing the objective in this form allows an intuitive interpretation for the semi-supervised training procedure. When supervision is provided, our objective function is evaluated at the observed value of z_s. When supervision is not provided, we evaluate the objective function for every possible value of z_s and take a (potentially infinite for continuous z_s) weighted average. The weighting in the average is given by $q(z_s|x, y, \phi)$, which is simultaneously trained to approximate the posterior $p(z_s|x, y, \theta)$. In other words, on unsupervised utterances, we evaluate our objective for each possible value of the latent attribute and weight by the (approximate) posterior probability that this value of the latent was responsible for generating the utterance.

As $q(z_s|x, y, \phi)$ is trained to approximate $p(z_s|x, y, \theta)$ we can expect it to become a reasonable classifier/regressor for the semi-supervised latent attribute as the model improves. However, this variational distribution is only trained on unsupervised training points and so does not benefit directly from the supervised data. To overcome this problem we follow Kingma & Welling (2013) and add a classification loss to our objective. The overall objective becomes:

$$\mathcal{L}_{total}(\theta, \phi) = \mathcal{L}(\theta, \phi) + \alpha E_{x,y,z_s}[\log q(z_s|x, y, \phi)] \tag{6}$$

Where α is a hyper parameter which adjusts the contribution of this term.

3 DATA

We have used a proprietary high quality labeled data-set of 40 English speakers. The training set consists of 72,405 utterances with durations of at most 5 seconds (52 hours). The validation and test sets each contain 745 utterances or roughly 30 minutes of data. We vary the amount of supervision in the experiments below. We also experimented with transferring controllability to a fully unlabeled data-set of audiobook recordings by Catherine Byers (the speaker from the 2013 Blizzard Challenge), which exhibits high variation in affect and prosody and to other speakers who were less expressive. We strongly encourage the reader to listen to the synthesized samples on our demo page.

[1] We define the differential entropy of the delta function to be 0

In this work we chose to focus on learning to control affect with a discrete representation, as well as speaking rate and F0 variation with a continuous representation, as these are challenging aspects of prosody to control. Our method could be applied to other factors without modification.

3.1 AFFECT CONTROL

The best way to represent emotion is an actively researched area and many models of affect exist. In this work we chose to follow the circumplex model of emotion (Russell, 1980) which posits that most affective states can be represented in a 2 dimensional plane with one axis representing arousal and the other axis representing valence. Arousal measures the level of excitement or energy and valence measures positivity or negativity. Figure 3, shows a chart of emotions plotted in the arousal-valence plane where we can see that, for example, high arousal and high valence corresponds to joy or happiness whereas high arousal and low valence might correspond to anger or frustration.

Our data-set was recorded under studio conditions with trained voice actors who were prompted to read dialogues in one of three valences:-2, -1, +2 and two arousal values:-2 (low), +2 (high). This was achieved by prompting the actors to read dialogues in either a happy, sad or angry voice at two levels of arousal. This results in 6 possible affective states which we chose to model as discrete and use as our supervision labels.

3.2 SPEAKING RATE AND F0 VARIATION CONTROL

In order to demonstrate that we can control continuous attributes we also created approximate real-valued labels for speaking rate and arousal for all of our data. We generate the approximate speaking rate as number of syllables per second in each utterance. F0, also known as the fundamental frequency, measures the frequency of vibration of the vocal folds during voiced sounds. Variation in F0 is highly correlated with arousal and roughly measures how expressive an utterance is. To create approximate arousal labels we extracted the F0 contour from each of our utterances, using the YIN algorithm (De Cheveigné & Kawahara, 2002), and measured its standard deviation. We then performed a whitening transform on these two approximate labels in order to match it to our standard normal prior.

These artificial labels would of course be cheap to obtain for the entire data-set and would not justify the use of semi-supervision in real applications. But, our objective here is to evaluate/demonstrate the efficacy of semi-supervision rather than to specifically control a particular attribute. We have chosen speaking rate and F0 variations, because they both correspond to subjecively distinct variations of interest, and they are more easily quantifiable than affect and so provide strong evidence of controllability. For the continuous latents we are not only able to interpolate speaking-rates and F0 variations but also to extrapolate outside of our training data. We provide examples on our demo page of samples with significantly greater/lower speed and variability than typically observed in natural speech.

4 EXPERIMENTS AND RESULTS

To evaluate the efficacy of semi-supervised latent variable models for controllable TTS we trained the model described in section 2 on the above data-sets at varying levels of supervision as well as for varying settings of the hyper-parameters: α which controls the supervision loss and γ, which over emphasizes supervised training points. We found that a value of $\alpha = 1$ was optimal for the discrete experiments and $\alpha = 0$ for the continuous experiments, which corresponds to simply optimizing the ELBO. We report results for varying levels of γ. $\gamma = 1$ corresponds to experiments with no over-weighting of the supervised points. All experiments were trained using the ADAM optimizer with learning rate of 10^{-3} and run for $300,000$ training steps with a batch size of 256, distributed across 32 Google Cloud TPU chips.

Assessing the degree of control is challenging as interpreting affect can be subjective. We used two objective metrics of control as well as subjective evaluation from human raters and a third objective metric to measure overall quality. For affect, the first objective metric we introduced was the test-set accuracy of a 6-class affect classifier trained on the ground truth training data and applied to generated samples from the model. The classifier is a convolutional neural network whose structure

Table 1: Mean Opinion Scores: Crowd sourced MOS for each of our models alongside 95% confidence intervals, showing no degradation in performance compared to the baseline.

			Semi-Supervised (10% supervision)		
			continuous latent		discrete latent
ground truth	baseline Tacotron	baseline with z_u	F0	speaking-rate	affect
4.521±0.073	4.088±0.094	4.244±0.079	4.276±0.072	4.158±0.081	4.167±0.086

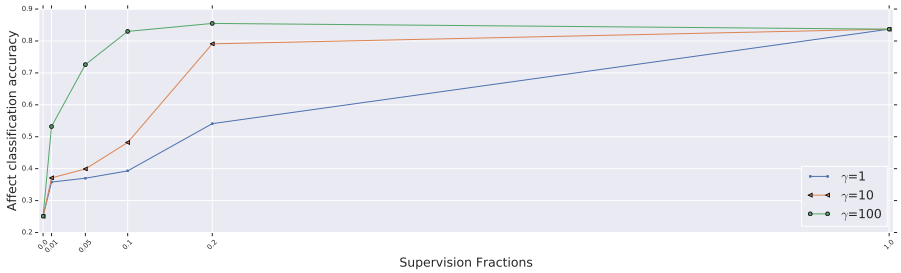
mirrors the posterior network $q(z_s|x, y, \phi)$ and its exact architecture is given in appendix A. Our personal subjective evaluation correlated highly with the classifier accuracy and we provide samples on our demo page. For speaking rate control, we are able to measure the syllable rate and so report the mean syllable rate error on a held out test-set. The syllable rate error is calculated as the absolute difference in syllable rate between the desired syllable rate and that measured from the synthesized sample. We calculate an analogous error rate for F0 variation.

Whilst the two metrics above measure controllability they don't tell us if this comes at the expense of a degradation in synthesis quality. To probe quality we use two further metrics. The first was Mel-Cepstral-Distortion-Dynamic-Time-Warping (MCD-DTW) (Kubichek, 1993) on a held out test-set. MCD-DTW is a measure of the difference between the ground-truth spectrogram and the synthesized mel spectrogram that is known to correlate well with human perception (Kubichek, 1993). The second metric of quality was crowd sourced mean-opinion-scores (MOS). Details of MOS and MCD calculations are provided in appendix B. We provide MOS figures at 10% supervision for all the models.

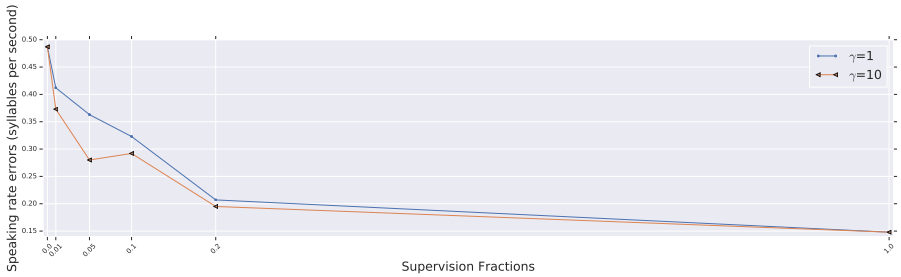
5 Discussion

The classification accuracy (see figure 4a) and error-rate results (see figure 4b-4c) provide a clear demonstration that using semi-supervised latent variables, we are able to achieve control of both continuous and discrete attributes of speech. There is not a significant degradation in the overall quality and this is evidenced by the mean opinion scores which are above the baseline, Tacotron. We also include a baseline of our Tacotron model augmented only by the unsupervised latent z_s, to aid comparison. The MCD-DTW scores for F0 variation and affect are improved at all levels of supervision (figure 4d). Whilst the MCD-DTW is degraded for speaking rate, this is likely a misleading metric when targeting changes in timing as the dynamic-time-warping component of MCD-DTW changes exactly the aspect we wish to control. For speaking rate the combination of MOS and samples is a better indication of the overall quality. We are able to reduce the supervision level to levels as low as 1% or 30 minutes and still have a significant degree of control. We show on our demo page that even at 15 minutes of supervision we can still achieve control of speaking rate and that we are able to extrapolate outside the range of values seen during training. On the affect data our classification accuracy doesn't degrade significantly until we reach 10% supervision and remains significantly above chance down to levels as low as 1% (30 minutes), see figure 4a. Obtaining 30 minutes of supervised data is likely within reach of most teams constructing TTS systems. Unlike previous work on generative modelling for control (Hsu et al., 2018; Wang et al., 2018), we do not require a post-processing stage to determine what our latent variables control and we can pre-determine what aspects we wish to control through choice of data. By separating our latent variables into those that are partially supervised and those which are fully unsupervised we retain the ability to model other latent aspects of prosody; this means that we can still draw samples of varying prosody whilst holding constant the affect or speaking rate.

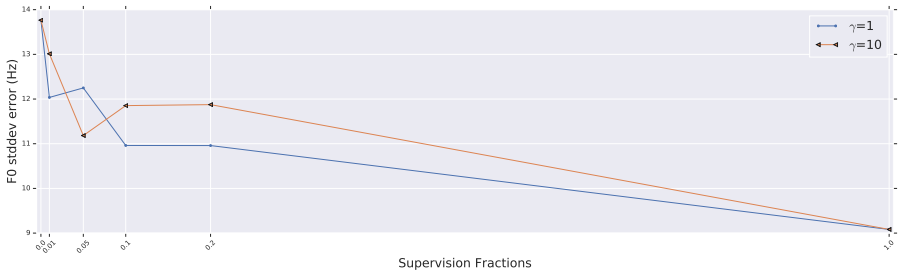
We observe the greatest degree of affect control, as measured by classifier accuracy, when $\alpha = 1$ and $\gamma = 100$. This means that to achieve the highest controllability we needed to 1) provide extra information to our approximate posterior $q(z_s|x, y, \phi)$ and 2) to over-represent the supervised data at low levels of supervision. Although both of these hyper-parameters have been used in the literature before (Narayanaswamy et al., 2017; Kingma et al., 2014) and shown to be either beneficial or necessary, they aren't strictly required by our probabilistic framework and so it is worth considering why they are needed. There are three potential sources of error in any generative model trained with SGVB: the model itself may be mis-specified such that the true data-generating distribution is not



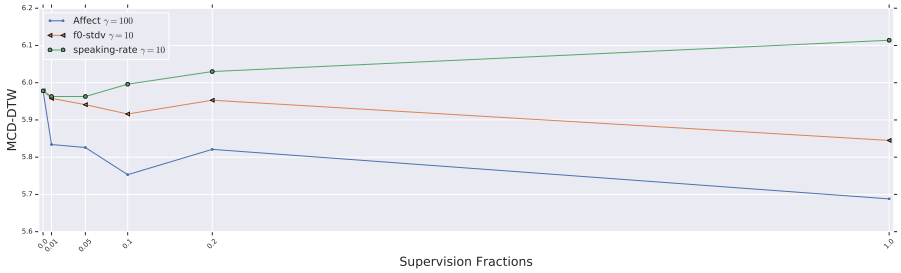
(a) Affect classification accuracy as a function of supervision level.



(b) Speaking-rate error as a function of supervision level.



(c) F0 variation error as a function of supervision level



(d) MCD-DTW as a function of supervision level for each of the three models. Lower is better and a supervision level of 0 corresponds to baseline Tacotron.

Figure 4: Objective evaluation metrics presented at multiple supervision levels.

in the model class, the parametric family chosen to approximate the posterior may be overly restrictive and finally the optimization landscape may contain undesirable local minima. These problems have afflicted previous work with deep latent variable models trained with SGVB, resulting in models that don't use their latent variables unless trained with complex annealing schedules (Bowman et al., 2015). In our case we believe that the necessity to set α and γ arises from a combination of model mis-specification and local minima. If α is set to 0, then at the start of training $q(z_s|x, y, \phi)$ is trained only to approximate $p(z_s|x, y, \theta_0)$, which is randomly initialized. We found empirically that in our discrete-latent experiments this resulted in $q(z_s|x, y, \phi)$ collapsing early in training to a point-mass on a single class for every single training example. Having ended up in this undesirable local minimum the posterior distribution never recovered, despite this being an obviously poor approximation to the model posterior later in training. The addition of the classification loss and

supervision weighting were sufficient to overcome this collapse and allow q to continue to model the posterior.

The optimization landscape is strongly affected by the relative size of the conditional likelihood and KL terms in our objective. These are in turn strongly affected by our choice of conditional independence assumptions and output-distributions. Thus, a natural direction for further work is to increase the expressivity of the conditional likelihood $p(x|y, z_s, z_u, \phi)$ to reduce model mis-specification. This could be done by learning the variance of the Laplace-distribution we currently use or by parameterizing more expressive output distributions that do not assume conditional independence across spectrogram channels. We conjecture that with more expressive output distributions, it may be possible to reduce the need for the α and γ terms in the objective. In this work we chose to use quite simple unconditional diagonal Gaussian priors, as our primary goal was to demonstrate the practicality of semi-supervision. Another natural extension would be to use conditional-priors $p(z|y)$ and to use more expressive priors such as mixtures as was done in Hsu et al. (2018).

5.1 Related Work

There has been enormous recent progress in neural TTS with numerous novel models proposed in recent years to synthesize speech directly from characters or phonemes (Shen et al., 2018; Arik et al., 2017; Gibiansky et al., 2017; Ping et al., 2017; Vasquez & Lewis, 2019; Taigman et al., 2017). Differentiating factors between these models include the degree of parallelism, with some models using Transformer based architectures (Ren et al., 2019), the choice of conditional independence assumptions made (Vasquez & Lewis, 2019) or the number of separately trained components (Gibiansky et al., 2017). Our work here is largely orthogonal to the exact structure of the conditional likelihood $P(x|y, z_s, z_u)$ and could be combined with all of the above methods.

Much of the recent research focus has been on modeling latent aspects of prosody. Early attempts include Global Style Tokens (Wang et al., 2018) which attempted to learn a trainable set of style-embeddings. Wang et al. (2018) condition the Tacotron decoder on a linear combination of embedding vectors whose weights during training are predicted from the ground-truth spectrogram. They were able to achieve prosodic control but there is no straightforward way to sample utterances of varying prosody. More recently, attempts have also been made to combine probabilistic latent variable models trained using SGVB (Akuzawa et al., 2018; Wan et al., 2019). These models use a fully unsupervised and non-identifiable approach, which makes it difficult to disentangle or interpret their latent variables for control. Hsu et al. (2018) attempt to overcome this problem by using a Gaussian mixture as the latent prior and so perform clustering in the latent space. Battenberg et al. (2019) introduce a hierarchical latent variable model to separate the modelling of style from prosody. However, all of these methods are fully unsupervised and this results in latents that can be hard to interpret or require complex post-processing.

The work most similar to ours is Wu et al. (2019) which also attempts to achieve affect control using semi-supervision with a heuristic approach based on Global Style Tokens (Wang et al., 2018). Wu et al. (2019) add a cross-entropy objective to the weightings of the style-tokens that encourages them to be one-hot on points with supervision. Similar to our method, they are able to achieve control over affect but unlike our method they do not have a principled probabilistic interpretation nor the ability to simultaneously model aspects of prosody other than emotion. The result is that their method is not able to draw samples of varying prosody for the same utterance with fixed emotion. Furthermore, whilst our method can be applied to both continuous and discrete controllable factors, its not clear how to extend the style-token based approach to handle continuous latent factors.

In the wider generative modelling literature, the combination of semi-supervision and deep latent variable models was first introduced in Kingma et al. (2014) who focus on using unlabelled data to improve classification accuracy. The potential to use the same technique for controllable generation was recognized by Narayanaswamy et al. (2017) who also provided demonstrations on image synthesis tasks. Since that work, interest in learning disentangled latent variables has grown but generally pursued alternate directions such as re-weighting the ELBO (Higgins et al., 2017), augmenting the objective to encourage factorization (Kim & Mnih, 2018) or using adversarial training (Mathieu et al., 2016). The ability to transfer controllability to speakers for whom we do not have supervision is referred to as domain transfer and our model bears similarities to that introduced by Ilse et al. (2019) but they use a mixture in their latent space more similar to Hsu et al. (2018).

5.2 ETHICAL CONSIDERATIONS

As with many advances in speech synthesis, progress in controllability raises the prospect that bad actors may misuse the technology either for misinformation or to commit fraud. Improvements in data efficiency and realism increase these risks and, when publishing, a consideration has to be made as to whether the benefits of the developments outweigh the risks. It is the opinion of the authors in this case that, since the focus of this work is on improved prosody, with potential benefits to human-computer interfaces, the benefits likely outweigh the risks. We nonetheless urge the research community to take seriously the potential for misuse both of this work and broader advances in TTS.

6 CONCLUSION

We have shown that the combination of semi-supervised latent variable models with neural TTS presents a practical and principled path towards building speech synthesizers we can control. Unlike previous fully unsupervised methods, we are able to consistently and reliably learn to control predetermined aspects of prosody. Our method can be applied to any latent attribute of speech for which a modest amount of labelling can be obtained, whether it be continuous or discrete. In our experiments we found that 30 minutes of supervision was sufficient, a volume of data that is within the reach of most research teams. We are able to learn to control subtle characteristics of speech such as affect and for continuous attributes we have provided demonstrations of extrapolation to ranges never seen during training, and to speakers with no supervision. Augmenting existing state-of-the-art TTS systems with latent variables does not degrade synthesis quality and we evidence this with crowd sourced mean opinion scores. Unlike similar heuristic methods, our probabilistic formulation, allows us to draw samples of varying prosody whilst holding constant some attribute we wish to control.

REFERENCES

K. Akuzawa, Y. Iwasawa, and Y. Matsuo. Expressive speech synthesis via modeling expressions with variational autoencoder. 2018.

S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, and J. Raiman. Deep voice: Real-time neural text-to-speech. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

E. Battenberg, S. Mariooryad, D. Stanton, R. Skerry-Ryan, M. Shannon, D. Kao, and T. Bagby. Effective use of variational embedding capacity in expressive end-to-end speech synthesis. *arXiv preprint arXiv:1906.03402*, 2019.

S. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

A. De Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, pp. 1917–1930, 2002.

A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou. Deep voice 2: Multi-speaker neural text-to-speech. In *Advances in neural information processing systems*, pp. 2962–2970, 2017.

A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2017.

W. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, et al. Hierarchical generative modeling for controllable speech synthesis. *International Conference On Learning Representations*, 2018.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 1999.

M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling. Diva: Domain invariant variational autoencoders. *arXiv preprint arXiv:1905.10427*, 2019.

N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Van den Oord, S. Dieleman, and K. Kavukcuoglu. Efficient neural audio synthesis. *arXiv preprint arXiv:1802.08435*, 2018.

H. Kim and A. Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pp. 3581–3589, 2014.

R Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1993.

M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pp. 5040–5048, 2016.

E. Munoz-de Escalona and J. J. Canas. Online measuring of available resources. 2017.

S. Narayanaswamy, B. T. Paige, J. Van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, and P. Torr. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, pp. 5925–5935, 2017.

W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*, 2017.

Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu. Fastspeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*, 2019.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The International Conference on Machine Learning*, 2014.

J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 1980.

M. Schröder. Emotional speech synthesis: A review. In *Seventh European Conference on Speech Communication and Technology*, 2001.

J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, and R. Skerrv-Ryan. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani. Voiceloop: Voice fitting and synthesis via a phonological loop. *arXiv preprint arXiv:1707.06588*, 2017.

A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

S. Vasquez and M. Lewis. Melnet: A generative model for audio in the frequency domain. *arXiv preprint arXiv:1906.01083*, 2019.

V. Wan, C. Chan, T. Kenter, J. Vit, and R. Clark. Chive: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network. *arXiv preprint arXiv:1905.07195*, 2019.

Y. Wang, RJ. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F Ren, Y. Jia, and R. A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*, 2018.

P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai. End-to-end emotional speech synthesis using style tokens and semi-supervised training. *arXiv preprint arXiv:1906.10859*, 2019.

A NEURAL NETWORK ARCHITECTURE

Sequence-to-Sequence model Our sequence-to-sequence network is modelled on Tacotron (Wang et al., 2018) but uses some modifications introduced in Skerry-Ryan et al. (2018). Input to the model consists of sequences of phonemes produced by a text normalization pipeline rather than character inputs. The CBHG text encoder from Wang et al. (2017) is used to convert the input phonemes into a sequence of text embeddings. The phoneme inputs are converted to learned 256-dimensional embeddings and passed through a pre-net composed of two fully connected ReLU layers (with 256 and 128 units, respectively), with dropout of 0.5 applied to the output of each layer, before being fed to the encoder. For multi-speaker models, a learned embedding for the target speaker is broadcast-concatenated to the output of the text encoder. The attention module uses a single LSTM layer with 256 units and zoneout of 0.1 followed by an MLP with 128 tanh hidden units to compute parameters for the monotonic 5-component GMM attention window. Instead of using the exponential function to compute the shift and scale parameters of the GMM components as in Graves (2013), we use the softplus function, which we found leads to faster alignment and more stable optimization. The attention weights predicted by the attention network are used to compute a weighted sum of output of the text encoder, producing a context vector. The context vector is concatenated with the output of the attention LSTM layer before being passed to the first decoder LSTM layer. The autoregressive decoder module consists of 2 LSTM layers each with 256 units, zoneout of 0.1, and residual connections between the layers. The spectrogram output is produced using a linear layer on top of the 2 LSTM layers, and we use a reduction factor of 2, meaning we predict two spectrogram frames for each decoder step. The decoder is fed the last frame of its most recent prediction (or the previous ground truth frame during training) and the current context as computed by the attention module. Before being fed to the decoder, the previous prediction is passed through a pre-net with the same same structure used before the text encoder above but its own paramters.

Variational Posteriors The variational distributions $q(z_s|x, y)$ and $q(z_u|x, y, z_s)$ are both structured as diagonal Gaussian distributions whose mean and variance are parameterized by neural networks. For discrete supervision we replace $q(z_s|x, y)$ by a categorical distribution and use the same network to output just the mean. The input to the distribution starts from the mel spectrogram x and passes it through a stack of 6 convolutional layers, each using ReLU non-linearities, 3x3 filters, 2x2 stride, and batch normalization. The 6 layers have 32, 32, 64, 64, 128, and 128 filters, respectively. The output of this convolution stack is fed into a unidirectional LSTM with 128 units. We pass the final output of this LSTM (and potentially vectors describing the text and/or speaker) through an MLP with 128 tanh hidden units to produce the parameters of the diagonal Gaussian posterior which we sample from. All but the last linear layer of these networks is shared between the two distributions $q(z_s|x, y)$ and $q(z_u|x, y, z_s)$. The resulting sample is broadcast-concatenated to the output of the text encoder. In our experiments z_u is always 32-dimensional and z_s is either a one-hot vector across 6 classes or a 1 dimensional continuous value.

Conditional inputs When providing information about the text to the variational posterior, we pass the sequence of text embeddings produced by the text encoder to a unidirectional RNN with 128 units and use its final output as a fixed-length text summary that is passed to the posterior MLP. Speaker information is passed to the posterior MLP via a learned speaker embedding.

WaveRNN We used the WaveRNN model described in Kalchbrenner et al. (2018) as our vocoder. We trained the network to map from synthesised mel-spectrograms to waveforms.

B EVALUATION

mel spectrograms The mel spectrograms the model predicts are computed from 24 kHz audio using a frame size of 50 ms, a hop size of 12.5 ms, an FFT size of 2048, and a Hann window. From the FFT energies, we compute 80 mel bins distributed between 80 Hz and 12 kHz.

MCD-DTW To compute mel cepstral distortion (MCD) (Kubichek, 1993), we use the same mel spectrogram parameters described above and take the discrete-cosine-transform to compute the first 13 MFCCs (not including the 0th coefficient). The MCD between two frames is the Euclidean distance between their MFCC vectors. Then we use the dynamic time warping (DTW) algorithm (with

a warp penalty of 1.0) to find an alignment between two spectrograms that produces the minimum MCD cost (including the total warp penalty). We report the average per-frame MCD-DTW.

Affect Classifier The affect classifier has a very similar structure to the variational posterior. The input to the classifier starts from the mel spectrogram x and passes it through a stack of 6 convolutional layers, each using ReLU non-linearities, 3x3 filters, 2x2 stride, and batch normalization. The 6 layers have 32, 32, 64, 64, 128, and 128 filters, respectively. The output of this convolution stack is fed into a unidirectional LSTM with 128 units. The final output of the LSTM is then passed through a softmax non-linearity to get logits over the training classes.

Mean Opinion Scores A human rater is presented with a single speech sample and is asked to rate perceived naturalness on a scale of 1 to 5, where 1 is "Bad" and 5 is "Excellent". For each sample, we collect 1 rating, and no rater is used for more than 6 items in a single evaluation. To analyze the data from these subjective tests, we average the scores and compute 95% confidence intervals. Natural human speech is typically rated around 4.5. Samples used for MOS from our model were drawn using the mean of z_u, whilst sampling z_s.