10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Using generative modelling to produce varied intonation for speech synthesis

Zack Hodari, Oliver Watts, Simon King

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
{zack.hodari, oliver.watts, Simon.King}@ed.ac.uk

Abstract

Unlike human speakers, typical text-to-speech (TTS) systems are unable to produce multiple distinct renditions of a given sentence. This has previously been addressed by adding explicit external control. In contrast, generative models are able to capture a distribution over multiple renditions and thus produce varied renditions using sampling.

Typical neural TTS models learn the average of the data because they minimise mean squared error. In the context of prosody, taking the average produces flatter, more boring speech: an "average prosody". A generative model that can synthesise multiple prosodies will, by design, not model average prosody.

We use variational autoencoders (VAEs) which explicitly place the most "average" data close to the mean of the Gaussian prior. We propose that by moving towards the tails of the prior distribution, the model will transition towards generating more idiosyncratic, varied renditions.

Focusing here on intonation, we investigate the trade-off between naturalness and intonation variation and find that typical acoustic models can either be natural, or varied, but not both. However, sampling from the tails of the VAE prior produces much more varied intonation than the traditional approaches, whilst maintaining the same level of naturalness.

Index Terms: speech synthesis, intonation modelling, prosodic variation, variational autoencoder, mixture density network

1. Introduction

Prosody in natural human speech varies predictably based on contextual factors. However, it also varies arbitrarily, or due to unknown factors [1]. Text-to-speech (TTS) voices are typically designed to synthesise a single most likely rendition of a given sentence. While many methods have been proposed to add control to TTS voices, often they do not take this arbitrary variation into account. In contrast, we focus on designing TTS voices that are able to produce any viable prosodic realisation of a given sentence in isolation. Such a system could be driven by contextual information (e.g. provided by a dialogue system) to produce more appropriate prosodic renditions. However, we here focus on the task of producing random (but acceptable) prosodic renditions given an isolated sentence.

Since neural statistical parametric speech synthesis (SPSS) became the leading paradigm in speech synthesis research [2] most TTS voices have used static plus dynamic features optimised with mean squared error, followed by maximum likelihood parameter generation (MLPG) and post-filtering [3]. These methods are a legacy of hidden Markov model (HMM) SPSS [4], where the problem of oversmoothing was observed and methods were developed to mitigate it. Oversmoothing of acoustic features is still an issue in neural SPSS, due to a combination of assumptions made in designing models [5]. Here we focus on prosody (and specifically on modelling intonation, which is the F_0 component of prosody) where the symptom of oversmoothing is flatter, more average prosody.

We argue that a model designed to synthesise distinct renditions will, by design, *not* model average prosody. Variational autoencoders (VAEs) are a class of generative models that can learn a smooth latent space approximating the true latent factors of the data. Therefore, we use a VAE [6] to tackle the problem of average prosody, using the latent space to capture otherwise unaccounted-for variation. We propose that by sampling from the low-probability regions of the VAE's prior we can generate idiosyncratic prosodic renditions.

2. Related work

Methods for control of SPSS voices roughly fall into two categories: explicitly labelled control and latent control. The former is typically expensive because labelling is labour-intensive, although this can be automated at the expense of accuracy [7, 8]. Labelling requires a concrete and consistent schema that can be followed by human annotators. For many aspects of variation in speech this is challenging, a clear example being emotion labelling [9]. For example, categorical emotions (e.g. happy or sad) may be too coarse, and appraisal-based measures (e.g. arousal or valence) may be too complex or ambiguous for labellers [10]. Additionally, there is the question of elicitation: should natural speech be annotated, or should the variation of interest be elicited (e.g., acted) and assumed to be correct?

It has been shown that unsupervised methods can achieve similar results to supervised control [11], which may be related to the challenge of accurately labelling variation in real data, as discussed above.

Discriminant condition codes, first proposed for speech recognition [12] have proved useful for multi-speaker TTS [13]. The same method has been applied in an unsupervised fashion [14], allowing for control of arbitrary variation. While these methods have been shown to have a probabilistic interpretation [11], they do not model uncertainty or guarantee smoothness in the latent space. As we discuss in Section 4, this smoothness is important for determining what corresponds to an idiosyncratic (and thus more varied) rendition of a sentence.

Tacotron [15] is a sequence-to-sequence model, for which style control using "global style tokens" (GST) has been proposed [16]. GSTs produce high quality speech, and can be predicted from text [17]. However, individual GSTs cannot be effectively used to produce distinct styles as they are trained as weighted combinations; using individual GSTs leads to significantly degraded audio quality. We expect a random weighting of the tokens will also produce degraded naturalness, since there is no smoothness constraint.

VAEs have been demonstrated for speech synthesis [18, 19], voice conversion [20], and intonation modelling [21, Chapter 7]. Discrete representations have also been incorporated into the VAE framework [22, 23]. An experiment with VQ-VAE [23] demonstrated that phones can be learnt with unsuper-

10.21437/SSW.2019-43

vised training, a result promising for potentially learning discrete prosodic styles. However, in this work we use a continuous latent space.

The recently-introduced clockwork hierarchical VAE (CHiVE) [24] is similar to the model we propose here, however our VAE does not make use of the clockwork hierarchical structure and we only predict intonation, while CHiVE predicts F_0, duration, and C_0. Since we consider isolated sentences, we are not concerned with a single "best" output of our system.

Prior work using VAEs has focused on modelling segmental features [23, 18], with some applications to intonation modelling, e.g. for style transfer [24] and predicting latents from text [21, Chapter 7]. However, our method moves towards TTS systems that can synthesise multiple distinct prosodic renditions (in an unsupervised framework and without the need for control).

3. Average prosody

While many methods have been proposed to add control, there is a more fundamental issue, known as oversmoothing, which leads to flatter, more boring prosody. Typical SPSS uses either feedforward neural networks, or recurrent neural networks (RNNs) to map from a linguistic specification to acoustic features. This mapping is learnt by minimising the mean squared error (MSE) against the ground truth acoustics. MSE is equivalent to minimising the negative log-likelihood (NLL) of a unit-variance Gaussian. This has two effects on such SPSS models: they learn the mean of the data, and are sensitive to outliers. By modelling the mean, SPSS models over-smooth the acoustics – in the context of prosody this is known as *average prosody*.

Methods such as the ϵ-contaminated Gaussian [25] exist to handle outliers. However, to fix both issues, it is common to collect speech that is as controlled and consistent as possible in terms of style. Training data with a single style results in models which produce more natural speech [26], but it also limits the voice's stylistic range. If we are interested in producing more varied style/prosody/intonation we need more varied data, but this must then be handled appropriately by our model.

Generative models, such as Mixture density networks (MDN) [27], have the ability to handle multiple modes. MDNs parameterise a Gaussian mixture model (GMM) for each acoustic frame which can help with oversmoothing of spectral features [28]. However, for prosodic features, we are interested in fixing oversmoothing over a longer timescale, for which frame-level GMMs are less suitable. Instead, we use variational autoencoders which model a distribution in an abstract (latent) space at whichever timescale is preferred.

4. Variational autoencoders

Variational autoencoders (VAEs) [6] are a class of latent variable models, i.e. they learn some unsupervised latent representation of the data. They consist of an encoder and a decoder: the encoder parameterises the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$, which is an approximation of $p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$ – the underlying factors that describe the data. The decoder is trained to reconstruct the input signal $\mathbf{x}$ from this latent space, i.e. given a sample from the posterior $\tilde{\mathbf{z}} \sim q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$, we reconstruct $\bar{\mathbf{x}} \sim p_{\boldsymbol{\theta}}(\mathbf{x} \mid \tilde{\mathbf{z}})$. The encoder and decoder are trained jointly by maximising the evidence lower bound (ELBO),

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) = &-KL(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \mid\mid p(\mathbf{z})) \\ &+ \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}\mid\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})\right] \qquad (1)\end{aligned}$$

The first term in the ELBO enforces a prior on the approximate posterior, while the second term measures reconstruction error. The Kullback-Leibler (KL) divergence term – used to enforce the prior – puts a cost on using the latent space. This cost on transmitting information through the latent space can encourage the approximate posterior to collapse to the prior, thus encoding no information: posterior collapse. KL-cost annealing is a common way to mitigate posterior collapse [29], where the KL term is down-weighted at the start of training, reducing the cost of encoding information in the latent space.

Here we consider conditional VAEs [30], which model F_0 conditioned on linguistic features. We use a sentence-level approximate posterior, although a sequence of phrase- or syllable-level latents would be a reasonable alternative. We use an isotropic Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{1})$, which gives an analytical form of the KL term.

Enforcing a Gaussian prior gives another useful quality: the single mode and smooth pdf means the distance of $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ from the prior mean $\mathbf{0}$ will be inversely proportional to the similarity of $\mathbf{x}$ and the largest mode in the data (e.g., the most common prosodic style). That is, the most idiosyncratic $\mathbf{x}$ will be far from the peak at $\mathbf{0}$. This is helpful for our interest in varied prosodic renditions; we can generate varied prosodic renditions using the decoder by sampling low-density regions in the prior. Thus we define two models that use only the VAE decoder,

$$\mathbf{z}_{\text{PEAK}} = \mathbf{0} \qquad \bar{\mathbf{x}}_{\text{PEAK}} \sim p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}_{\text{PEAK}}) \qquad (2)$$

$$\mathbf{z}_{\text{TAIL}} \sim vMF(\kappa = 0) \qquad \bar{\mathbf{x}}_{\text{TAIL}(r)} \sim p_{\boldsymbol{\theta}}(\mathbf{x} \mid r \times \mathbf{z}_{\text{TAIL}}) \qquad (3)$$

where $\bar{\mathbf{x}}_{\text{PEAK}}$ should correspond to the most common mode, i.e. style. Due to the uni-modal prior $p(\mathbf{x})$, $\bar{\mathbf{x}}_{\text{PEAK}}$ may instead correspond to an average of multiple styles, i.e. average prosody. Our proposed model uses $\mathbf{z}_{\text{TAIL}}$ (uniform samples on a hypersphere's surface[1]) to produce idiosyncratic renditions $\bar{\mathbf{x}}_{\text{TAIL}(r)}$, where the larger the radius r is the more unlikely the rendition.

5. Systems

We focus on modelling intonation, though in the future we plan to extend this to complete prosodic modelling (F_0, duration and energy). Modelling only F_0 limits the range of variation we can achieve, but reduces the risk of producing unnatural speech: spectral features and durations are taken from natural speech in our experiments, with full TTS left for future work. We use the WORLD vocoder [31], for analysis and synthesis. Our models[2] are implemented in PyTorch [32].

We use the same basic recurrent architecture for all trainable modules in Figure 1: a feedforward layer with 256 units, followed by three uni-directional recurrent layers using gated recurrent cells (GRUs) [33] with 64 units, finally any outputs used are projected to the required output dimension.

We use 600-dimensional linguistic labels from the standard Unilex question-set and 9 frame-level positional features with min-max normalisation as in the standard Merlin recipe [34]. The model predicts $\log F_0$, delta (velocity), and delta-delta (acceleration) features with mean-variance normalisation. We use Adam [35] with an initial learning of 0.005, which is increased linearly for the first 1000 batches, and then decayed proportional to the inverse square of the number of batches [36, Sec 5.3], where our batch size is 32. Early stopping is used

[1] Sampled from a von Mises-Fisher distribution (vMF) with uniform concentration – wikipedia.org/wiki/Von_MisesFisher_distribution

[2] Code is available at github.com/ZackHodari/average_prosody

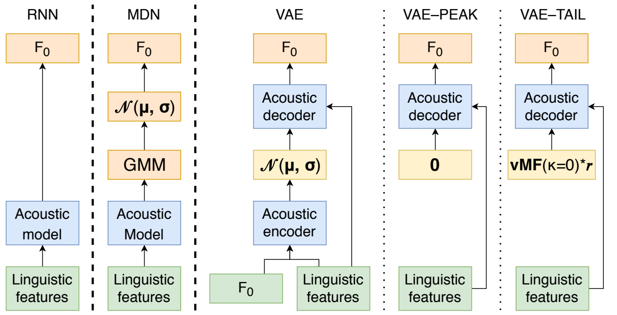

Figure 1: *Illustration of our models, where only the first three are trained models.* VAE–PEAK *and* VAE–TAIL *are different configurations of the VAE model. Blue: learned modules. Green: frame-level inputs. Orange: frame-level predictions. Yellow: sentence-level latent space.*

based on validation performance. MLPG [37] is used to generate the F_0 contour from the dynamic features; predicted standard deviations are used by the MDN, and all other models use the global standard deviation of the training data.

The MDN uses four mixture components, whose variances are floored at 10^{-4}. To synthesise from the MDN, we use the most likely component sequence (i.e. argmax) to select means and variances used in MLPG.

Systems VAE–PEAK and VAE–TAIL in the list below are identical apart from the use of different sampling schemes (see Section 4). Their shared model uses a 16-dimensional isotropic Gaussian as the approximate posterior. The latent sample $\tilde{\mathbf{z}}$ is broadcast to frame-level and input to the decoder, along with the linguistic features. The decoder predicts static and dynamic $\log F_0$ features; as such the reconstruction loss is MSE. The KL-divergence term is weighted by zero during the first epoch and increased linearly to 0.01 over 40 epochs. Using this annealing schedule, the model converged to a KL-divergence of 3.13.

RNN	Standard RNN-based SPSS model, using MSE.
MDN	MDN with 4 mixture components, using NLL.
VAE–PEAK	VAE decoder using $\mathbf{z}_{\text{PEAK}}$, i.e. the zero vector.
VAE–TAIL	VAE decoder using $\mathbf{z}_{\text{TAIL}}$ with $r = 3$, i.e. points on the surface of a hypersphere with radius 3.
COPY–SYNTH	Natural F0.
BASELINE	A quadratic polynomial fitted to natural F_0.
RNN–SCALED	F_0 from RNN, scaled vertically by a factor of 3.

5.1. Purpose of baselines

BASELINE sets a lower bound on naturalness (and variedness): no matter how much variation a system produces, its naturalness should never fall below that of BASELINE. An upper-bound is COPY–SYNTH: no system should be more natural than this, but might sound more varied even though it is unclear whether this would be favoured by listeners.

Because we expect that adding more variation will degrade naturalness, we wish to quantify this. RNN–SCALED is intended as a lower-bound on naturalness using a naïve method for increasing variation, similar to variance scaling [38]. RNN–SCALED is intended to demonstrate that VAE–TAIL can produce the same amount of perceived variation but *without sacrificing* as much naturalness. In this study, setting the amount of perceived variation in RNN–SCALED and VAE–TAIL was calibrated in a pilot listening test by the authors, where we attempted to match the level of variation to COPY–SYNTH.

6. Hypotheses

$\mathbf{H_1}$ VAE–TAIL will be much more **varied** than the typical SPSS systems (RNN, VAE–PEAK, MDN).

$\mathbf{H_2}$ RNN–SCALED, VAE–TAIL, and COPY–SYNTH will have the same level of **variedness**.

$\mathbf{H_3}$ RNN, VAE–PEAK, and MDN will have a similar level of **variedness**, where MDN is more varied than the other two.

$\mathbf{H_4}$ VAE–TAIL will have slightly lower **naturalness** than the typical SPSS systems (RNN, VAE–PEAK, MDN).

$\mathbf{H_5}$ VAE–TAIL will be much more **natural** than the varied baseline RNN–SCALED.

7. Data

Our choice of training data is motivated by the need for prosodic variation: if the data is very stylistically consistent, there will be too little variation for the VAE to capture in its latent space. We therefore use the Blizzard Challenge 2018 dataset [39] provided by Usborne Publishing. The data consists of stories read in an expressive style for a 4–6 year old audience, with some character voices. Many of the stories include substantial amounts of direct speech. In total it contains 6.5 hours (~7,250 sentences) of professionally-recorded speech from a female speaker of standard southern British English. The training-validation-test split described in Watts et al. [14] is used.

8. Evaluation

We want to evaluate the amount of variation produced by the systems described. However, variation alone is not a guarantee of "better" speech synthesis [40]. For this reason we evaluate quality along with variation. To determine quality we measure naturalness using a standard mean opinion score test, where users were asked to "rate the naturalness" on a 5-point Likert scale.

Evaluating variation is less straightforward. We employed a preference test where two systems were compared side by side for the same sentence. Users were asked to choose "which sentence has more varied intonation", where one sentence must be be marked as "more flat", and the other as "more varied". Due to the large number of pairs for 7 systems, we excluded BASELINE in the pairwise test, as it is clear from the speech samples[3] that it would be the least varied. However, without BASELINE in the variation test we lose our lower-bound on variation.

We randomly selected 32 test sentences of between 7 and 11 words (1.4 to 4.8 seconds). The naturalness test was performed before the preference test. As there were 22 screens to be completed for each sentence it was necessary to split the test into two halves using a simple 2x2 Latin square between-subjects design. In total we used 30 participants, 15 per listener group, the test took 45 minutes and participants were paid £8.

9. Results

9.1. Naturalness test

A summary of the naturalness ratings is provided in Figure 2. We perform a Wilcoxon rank-sums significance test between all pairs of systems in the naturalness test, followed by Holm-Bonferroni correction. This statistical analysis is the same as for the Blizzard challenge [41]. VAE–TAIL, RNN, MDN, and VAE–PEAK form a group for which we did not find any significant

[3] Speech samples available at github.com/ZackHodari/average_prosody

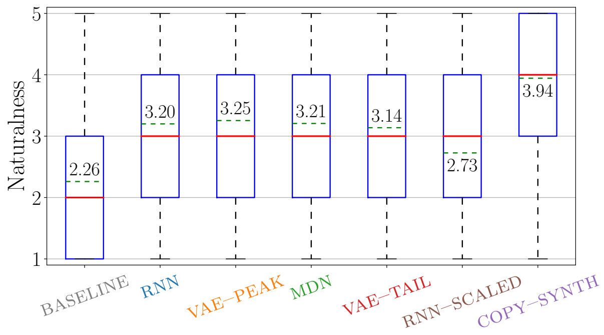

Figure 2: *Naturalness results. Solid red lines are medians, dashed green lines are means (cannot be used for statistical comparison), blue boxes show the 25th and 75th percentiles, and whiskers show the range of the ratings, excluding outliers which are plotted with +. Ordered according to the variation test.*

difference. All other system pairs are significantly different, with a corrected p-value of less than 0.00001.

9.2. Variation test

While it is not guaranteed that human preferences are self-consistent, or globally consistent[4], we see that the results in Figure 3 do form a consistent ordering from most flat to most varied: RNN → VAE–PEAK → MDN → VAE–TAIL → COPY–SYNTH → RNN–SCALED. However, relative variedness is sometimes inconsistent, e.g. while RNN–SCALED is more varied than COPY–SYNTH (5th row), we see that the difference between COPY–SYNTH and RNN (13th row) is greater than the difference between RNN–SCALED and RNN (15th row).

We perform a binomial significance test for the 15 pairs in the listening test, followed by Holm-Bonferroni correction. With the correction we find that (RNN, VAE–PEAK), (VAE–PEAK, MDN), and (COPY–SYNTH, RNN–SCALED) did not show a significant difference: this is indicated by the colouring of those pairs in Figure 3. All other pairs are significantly different, with a corrected p-value of less than 0.0002.

9.3. Naturalness–Variedness trade-off

While this ordering supports our expectations, we cannot clearly comment on their support of our hypotheses in Section 6 as the relative variedness between systems is not clear. Additionally, we would like to clearly compare the trade-off between increasing intonation variation and naturalness. This requires us to represent the pairwise preferences in Figure 3 along a single axis.

We could approach this using multi-dimensional scaling (MDS) [43]; however, the pairwise preferences correspond to directed edges, not distances. Instead, we formulate the problem as a system of linear equations[5]. Here, the variables are the positions of each system in the dimension of relative variedness, and each equation describes the "excess preference" of a system pair (the difference between the two system's average preference). This system can be solved using ordinary least squares:

$$Ax = b \qquad\qquad x = (A^T A)^{-1} A^T b$$

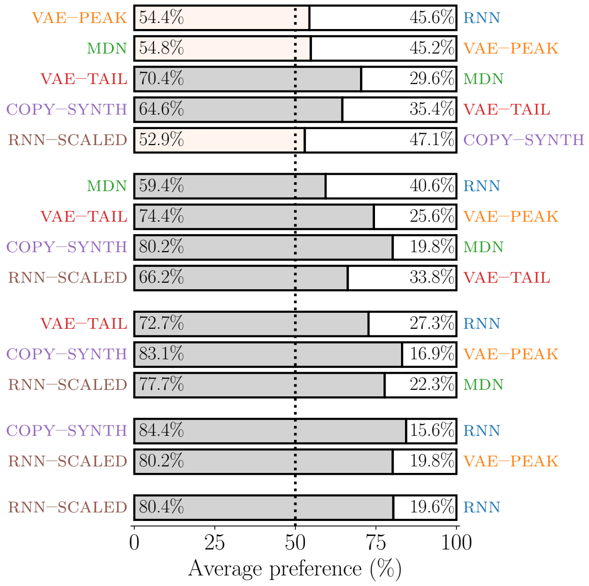

Figure 3: *Pairwise variedness results. Pairs are ordered such that the more varied system is on the left. The top 5 rows give the pairs that are consecutive in the ordering, with following rows showing systems that are increasingly further apart in the ordering. We did not find a significant difference for the pairs marked in a lighter colour.*

where $A \in \{-1, 0, 1\}^{15\times 6}$ and $b \in \mathbb{R}^{15\times 1}$ encode the pairwise results in Figure 3. Given the solution ($x \in \mathbb{R}^{6\times 1}$) we plot naturalness against relative variedness in Figure 5. Systems to the left have flatter intonation, and systems to the right have more varied intonation. This axis represents human preference and is not intended to be a perceptual scale.

In Figure 5, we see that VAE–TAIL is much more varied than the typical SPSS systems ($\mathbf{H_1}$). It is also clear that our calibration favoured less variation in VAE–TAIL than COPY–SYNTH (rejecting $\mathbf{H_2}$), thus we cannot make broad statements about the naturalness-variedness trade-off. However, based on the significant drop in naturalness from RNN to RNN–SCALED, and the clustering over relative variedness, we believe that VAE–TAIL would still be significantly more natural than RNN–SCALED even if it matched COPY–SYNTH's level of variation.

RNN, VAE–PEAK, and MDN are clustered along the axis of relative variation, with MDN being significantly more varied, but only by a small amount ($\mathbf{H_3}$). Demonstrating that all systems suffer from oversmoothing of F_0 to a similar extent.

While the mean naturalness of VAE–TAIL is lower than RNN, VAE–PEAK, and MDN, the means cannot be directly compared, and no significant difference was found in Section 9.1. Rejecting $\mathbf{H_4}$ suggests we can produce more varied intonation without sacrificing naturalness. However, we expect that with the ideal calibration we may see some slight degradation in naturalness of VAE–TAIL. We do observe that VAE–TAIL is much more natural than RNN–SCALED ($\mathbf{H_5}$).

10. Analysis

Calibration The horizontal axis in Figure 5 shows VAE–TAIL having much greater perceived intonation variation than MDN, while the logF_0 histograms in Figure 4 shows them as having

[4] As described by Arrow's impossibility theorem [42].

[5] We thank Erfan Loweimi and Gustav Henter for insightful discussions that led to this formulation of the problem.

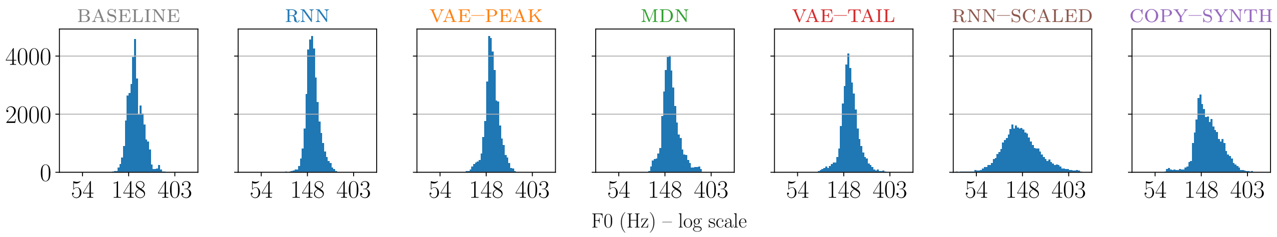

Figure 4: *Histogram of logF$_0$ values for each system over all the listening test material. Ordered according to the variation test.*

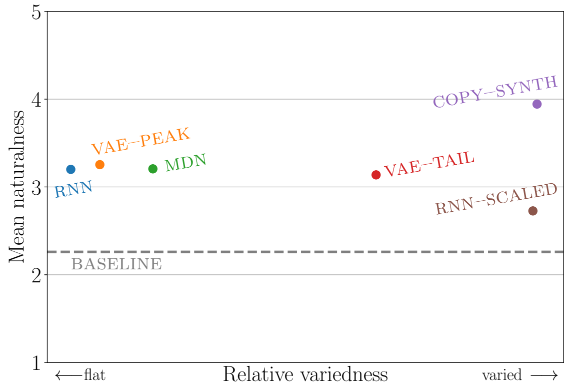

Figure 5: *Naturalness-variedness trade-off. Ideally as we increase the amount of prosodic variation our system will not decrease in naturalness. Note that naturalness comparisons can only be made using the significance results in Section 9.1.*

the same amount of objective variation – variance of logF$_0$ predictions for the listening test stimuli. This demonstrates that objective measures do not necessarily correspond to perceived variation, which is exactly what makes calibration of VAE–TAIL and RNN–SCALED difficult. Figure 4 shows that VAE–TAIL has a narrower histogram than COPY–SYNTH, however as objective measures do not necessarily correspond to perceived variation we chose not to rely on objective measures for calibration.

Multiple renditions We have demonstrated the ability to produce varied intonation while maintaining the same level of naturalness, thus mitigating average prosody. However, we have not demonstrated VAE–TAIL's ability to produce multiple distinct prosodic renditions. In Figure 6 we present a density of 10,000 F$_0$ contours $\bar{\mathbf{x}}_{\text{TAIL}(3)}$ produced using samples $\mathbf{z}_{\text{TAIL}} \sim vMF(\kappa = 0)$. As expected, the F$_0$ contours produced vary smoothly, but more importantly we see that they vary between multiple distinct contours. For this sentence we see that there may be three distinct contours. We are interested in evaluating the distinctiveness of multiple different samples from VAE–TAIL; however, this is out of the current scope.

MDN sampling While MDN is also a generative model, sampling from the frame-level GMMs is not straightforward. MLPG can be used to select the single best trajectory [37, Case 3]. But to produce multiple renditions from MDN we must choose a sequence of Gaussian components. However, randomly choosing components produces noisy F$_0$ contours, and using the same component for the entire sequence does not produce distinct performances. This is likely because the components don't represent modes of the data, but behave in a similar way to the ϵ-contaminated Gaussian distribution [25].

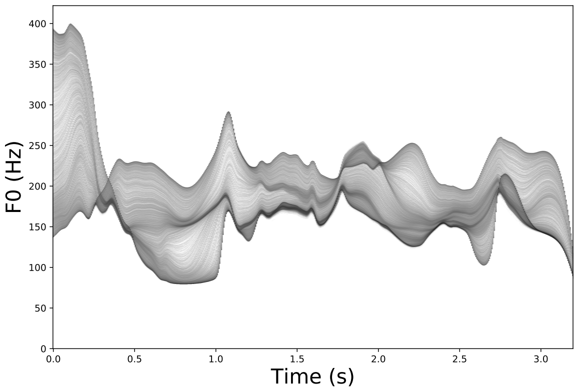

Figure 6: *Density of F$_0$ predictions made by* VAE–TAIL *for the sentence "Goldilocks skipped around a corner and saw..."*

11. Conclusion

We have demonstrated that output from typical RNN and MDN models exhibits flat intonation. Additionally, we have provided evidence that sampling from the tails of a VAE prior produces speech that is much more varied than typical SPSS while maintaining the same level of naturalness. In future we plan to undertake a full evaluation of this trade-off, to determine if and when this method begins to improve or degrade in quality.

In future work, we plan to: use MUSHRA in place of a preference test; use a neural vocoder; make use of seq2seq models with attention instead of upsampling the linguistic features; predict other prosodic features; and make use of either a discrete latent space [22] or a mixture model VAE prior [44].

Acknowledgements: Zack Hodari was supported by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh. Oliver Watts was supported by EPSRC Standard Research Grant EP/P011586/1.

12. References

[1] Y. Xu, "Speech prosody: A methodological review," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 85–115, 2011.

[2] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, p. 6, 2014.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*. Vancouver, Canada: IEEE, 2013, pp. 7962–7966.

[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[5] G. E. Henter, T. Merritt, M. Shannon, C. Mayo, and S. King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proc. Interspeech*, Singapore, 2014, pp. 1504–1508.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[7] A. Rosenberg, "AuToBI - a tool for automatic ToBI annotation," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 146–149.

[8] Z. Hodari, O. Watts, S. Ronanki, and S. King, "Learning interpretable control dimensions for speech synthesis by using external data," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 32–36.

[9] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 1, pp. 33–60, 2003.

[10] Z. Hodari, "A learned emotion space for emotion recognition and emotive speech synthesis," Master's thesis, The University of Edinburgh, 2017.

[11] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi, "Deep encoder-decoder models for unsupervised learning of controllable speech synthesis," *arXiv preprint arXiv:1807.11470*, 2018.

[12] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1713–1725, 2014.

[13] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in *Proc. ICASSP*. New Orleans, USA: IEEE, 2017, pp. 4905–4909.

[14] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 2217–2221.

[15] Y. Wang, R. 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 text-to-speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[16] Y. Wang, D. Stanton, Y. Zhang, R. 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.

[17] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," *arXiv preprint arXiv:1808.01410*, 2018.

[18] W.-N. Hsu, Y. Zhang, R. Weiss, H. Zen, Y. Wu, Y. Cao, and Y. Wang, "Hierarchical generative modeling for controllable speech synthesis," in *Proc. ICLR*, New Orleans, USA, 2019.

[19] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," *arXiv preprint arXiv:1804.02135*, 2018.

[20] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[21] X. Wang, "Fundamental frequency modelling for neural-network-based statistical parametric speech synthesis," Ph.D. dissertation, SOKENDAI – The Graduate University for Advanced Studies, 2018.

[22] J. T. Rolfe, "Discrete variational autoencoders," *arXiv preprint arXiv:1609.02200*, 2016.

[23] A. van den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Proc. NeurIPS*, Long Beach, USA, 2017, pp. 6306–6315.

[24] V. Wan, C.-a. Chan, T. Kenter, J. Vit, and R. Clark, "CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *Proc. ICML*, Long Beach, USA, 2019.

[25] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak, "Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices," in *Proc. Interspeech*, San Francisco, USA, 2016, pp. 2273–2277.

[26] M. Podsiadlo and V. Ungureanu, "Experiments with training corpora for statistical text-to-speech systems," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 2002–2006.

[27] C. M. Bishop, "Mixture density networks," Citeseer, Tech. Rep., 1994.

[28] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. ICASSP*. Florence, Italy: IEEE, 2014, pp. 3844–3848.

[29] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. ICLR*, vol. 3, Toulon, France, 2017.

[30] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Proc. NeurIPS*, Montreal, Canada, 2015, pp. 3483–3491.

[31] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS-W*, Long Beach, USA, 2017.

[33] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[34] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. Speech Synthesis Workshop*, Sunnyvale, USA, 2016, pp. 124–124.

[35] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, Long Beach, USA, Dec 2017, pp. 5998–6008.

[37] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 3. Istanbul, Turkey: IEEE, 2000, pp. 1315–1318.

[38] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *Proc. Interspeech*, Portland, USA, 2012, pp. 1436–1439.

[39] S. King, J. Crumlish, A. Martin, and L. Wihlborg, "The Blizzard challenge 2018," in *Proc. Blizzard Challenge Workshop*, Hyderabad, India, 2017.

[40] J. Latorre, K. Yanagisawa, V. Wan, B. Kolluru, and M. J. Gales, "Speech intonation for TTS: Study on evaluation methodology," in *Proc. Interspeech*, Singapore, 2014, pp. 2957–2961.

[41] R. A. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard challenge 2007 listening test results," in *Proc. Blizzard Challenge Workshop*, Bonn, Germany, 2007.

[42] K. J. Arrow, "A difficulty in the concept of social welfare," *J. of political economy*, vol. 58, no. 4, pp. 328–346, 1950.

[43] I. Borg and P. Groenen, "Modern multidimensional scaling: Theory and applications," *Journal of Educational Measurement*, vol. 40, no. 3, pp. 277–280, 2003.

[44] J. M. Tomczak and M. Welling, "VAE with a VampPrior," in *Proc. Artificial Intelligence and Statistics*, Lanzarote, Spain, 2018, pp. 1214–1223.