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 F0 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 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 labelers [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 5, 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], 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 phonemes can be learnt with unsuper-
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 F0, duration, and C0. 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 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 – they learn the mean of the data, and are sensitive to outliers. Thus we define two models that use only the VAE decoder, $z_{0} = 0$ $x_{0} \sim p(x, z_{0})$ $x_{tail} \sim vMF(\kappa = 0)$ $x_{tail(r)} \sim p(x, r \times x_{tail})$ where $x_{peak}$ should correspond to the most common mode, i.e. style. Due to the unimodal prior $p(x)$, $x_{peak}$ may instead correspond to an average of multiple styles, i.e. average prosody. Our proposed model uses $x_{tail}$ (uniform samples on a hypersphere’s surface1) to produce idiosyncratic renditions $x_{tail(r)}$, where the larger the radius $r$ is the more unlikely the rendition.

### 4. Variational autoencoders

Variational autoencoders (VAEs) [8] 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 $q_{φ}(z | x)$, which is an approximation of $p_{θ}(z | x)$ – the underlying factors that describe the data. The decoder is trained to reconstruct the input signal $x$ from this latent space, i.e. given a sample from the posterior $z \sim q_{φ}(z | x)$, we reconstruct $x \sim p_{θ}(x | z)$. The encoder and decoder are trained jointly by maximising the evidence lower bound (ELBO),

$$L(θ, φ; x) = -KL(q_{φ}(z | x) \parallel p(z)) + E_{q_{φ}(z|x)}[\log p_{θ}(x | z)]$$

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 $p(z) = N(0, \theta, 1)$, 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 $q_{φ}(z | x)$ from the prior mean $0$ will be inversely proportional to the similarity of $x$ and the largest mode in the data (e.g., the most common prosodic style). That is, the most idiosyncratic $x$ will be far from the peak at $0$. 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,

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

2. Code is available at [github.com/ZackHodari/average_prosody](https://github.com/ZackHodari/average_prosody)
based on validation performance. MLPG 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 \( \tilde{z} \) 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 \( z_{\text{PEAK}} \), i.e. the zero vector.

**VAE–TAIL** VAE decoder using \( z_{\text{TAIL}} \) with \( r = 3 \), i.e. points on the surface of a hypersphere with radius 3.

**COPY–SYNTH** Natural \( F_0 \).

**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

**H\(_1\)** VAE–TAIL will be much more varied than the typical SPSS systems (RNN, VAE–PEAK, MDN).

**H\(_2\)** RNN–SCALED, VAE–TAIL, and **COPY–SYNTH** will have the same level of variedness.

**H\(_3\)** RNN, VAE–PEAK, and MDN will have a similar level of variedness, where MDN is more varied than the other two.

**H\(_4\)** VAE–TAIL will have slightly lower naturalness than the typical SPSS systems (RNN, VAE–PEAK, MDN).

**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 [59] 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. [13] 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 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 [7] 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 [51]. VAE–TAIL, RNN, MDN, and VAE–PEAK form a group for which we did not find any significant

---

7Speech samples available at [github.com/ZackHodari/average_prosody](https://github.com/ZackHodari/average_prosody)
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 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 (15th 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 0.00002.

9.3. Naturalness–Variedness trade-off

While this ordering supports our expectations, we cannot clearly comment on their support of our hypotheses in Section 4 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. 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:

\[ Ax = b \]

\[ x = (A^T A)^{-1} A^T b \]

As described by Arrow’s impossibility theorem \(^{[43]}\).

We thank Erfan Losevim and Gustav Henter for insightful discussions that led to this formulation of the problem.

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 (H1). It is also clear that our calibration favoured less variation in VAE–TAIL than COPY–SYNTH (rejecting H2), 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 (H1). Demonstrating that all systems suffer from oversmoothing of F0 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 7.1. Rejecting H1 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 (H2).

10. Analysis

Calibration The horizontal axis in Figure 5 shows VAE–TAIL having much greater perceived intonation variation than MDN, while the logF0 histograms in Figure 4 shows them as having
the same amount of objective variation – variance of log$F_0$ 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 4 we present a density of 10,000 $F_0$ contours $x_{\text{TAIL}}(3)$ produced using samples $z_{\text{TAIL}} \sim \nu MF(\kappa = 0)$. 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 [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 $\epsilon$-contaminated Gaussian distribution [25].

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 Bayers,” 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, and S. Kambe, 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 model 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. Papadogiannis, 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. Chintal, 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, E. 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. Sielén, E. Helander, J. Nurminen, and M. Gabbo, “Towards better automatic speech recognition,” in Proc. ICASSP, Portland, USA, 2012, pp. 1436–1439.

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

[40] J. Latorre, K. Yangisawa, V. Wan, B. Kolluru, and M. J. Gales, “Speech intonation for TTS: Study on evaluation methodology,” in Proc. Interspeech, Bonn, Germany, 2015, pp. 2217–2221.

[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 Economic Theory and applications,” in Proc. Interspeech, Singapore, 2014, pp. 2957–2961.

[43] 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.

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

[45] 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.