Comparing normalizing flows and diffusion models for prosody and acoustic modelling in text-to-speech

Guangyan Zhang[1,2], *Thomas Merritt*[1], *Manuel Sam Ribeiro*[1], *Biel Tura-Vecino*[1], *Kayoko Yanagisawa*[1], *Kamil Pokora*[1], *Abdelhamid Ezzerg*[1], *Sebastian Cygert*[1,3], *Ammar Abbas*[1], *Piotr Bilinski*[1,4], *Roberto Barra-Chicote*[1], *Daniel Korzekwa*[1], *Jaime Lorenzo-Trueba*[1]

[1] Amazon TTS, [2] Department of Electronic Engineering, The Chinese University of Hong Kong, [3] Gdańsk University of Technology, [4] University of Warsaw

{manuerib, truebaj}@amazon.com

Abstract

Neural text-to-speech systems are often optimized on $\mathcal{L}1/\mathcal{L}2$ losses, which make strong assumptions about the distributions of the target data space. Aiming to improve those assumptions, Normalizing Flows and Diffusion Probabilistic Models were recently proposed as alternatives. In this paper, we compare traditional $\mathcal{L}1/\mathcal{L}2$-based approaches to diffusion and flow-based approaches for the tasks of prosody and mel-spectrogram prediction for text-to-speech synthesis. We use a prosody model to generate *log-f0* and duration features, which are used to condition an acoustic model that generates mel-spectrograms. Experimental results demonstrate that the flow-based model achieves the best performance for spectrogram prediction, improving over equivalent diffusion and $\mathcal{L}1$ models. Meanwhile, both diffusion and flow-based prosody predictors result in significant improvements over a typical $\mathcal{L}2$-trained prosody models.

Index Terms: text-to-speech, prosody modelling, acoustic model, normalizing flows, diffusion

1. Introduction

Neural text-to-speech (TTS) has recently demonstrated significant success in generating high-quality and stable speech [1–5]. However, TTS systems are still affected by the one-to-many mapping problem caused by speech containing many possible variations not directly explained by the phoneme sequence, such as prosody [2] or emotion [6]. The typical approach of training with $\mathcal{L}1$ or $\mathcal{L}2$ losses pushes the model to produce 'average' (over-smoothed) mel-spectrograms, resulting in synthesised speech with flat prosody and low quality [7]. Two strategies can be applied to handle this problem. On the one hand, we may provide auxiliary inputs to the acoustic model, such as explicit prosodic features [4, 5]. This has the additional advantage of allowing disentangled prosody control, or transfer [8]. Alternatively, we may move from a point-based estimation approach (e.g. models using $\mathcal{L}1$ or $\mathcal{L}2$ losses) to a probability density estimation approach (models using Normalizing Flows [9, 10] or Diffusion Models [11, 12]). We compare both strategies and analyze their impact on text-to-speech synthesis.

In terms of the first strategy, we may use two acoustic correlates of prosody, duration and $f0$, to explain speech variation. This helps the model to disambiguate the one-to-many problem of TTS synthesis. Oracle prosodic features extracted from the target speech can be provided to the model during training. However, for inference a prosody predictor is required to provide the prosodic features to the acoustic model. In previous work [4, 5, 13], the $\mathcal{L}2$ loss function is applied to optimize the prosody predictors. However, this loss results in the predictor generating average prosody, which lacks expressivity. In [14], a Mixture Density Network (MDN), is used to overcome the prosody prediction over-smoothing problem. However, the authors investigate only $f0$, and, for acoustic modelling, rely on an older Statistical Parametric Speech Synthesis (SPSS) system, rather than a more recent neural TTS approach. More recently, [15] found encouraging results when applying normalizing flows to the task of duration modelling. However, this prediction was independent of $f0$. In this paper, state-of-art generative models, normalizing flow and diffusion, are investigated for the task of joint $f0$ and duration modelling.

It is the *goal of this paper* to compare a traditional $\mathcal{L}1/\mathcal{L}2$-loss approach to probability density estimation approaches using Normalizing Flows and Diffusion Models. We investigate these architectures for the tasks of prosody modelling (generation of *f0* and duration) and acoustic modelling (generation of mel-spectrograms). Our key contributions are: 1) the use normalizing flows and diffusion models to address the problem of over-smoothing for joint $f0$ and duration prediction; and 2) a direct comparison of typical $\mathcal{L}1/\mathcal{L}2$-loss based approaches to normalising flows and diffusion models for the tasks of acoustic and prosody modelling.

2. Models

We investigate normalizing flows and diffusion models for the tasks of prosody and acoustic modelling. These are compared against standard $\mathcal{L}1/\mathcal{L}2$-based approaches from literature.

$\mathcal{L}1/\mathcal{L}2$ loss is commonly applied in literature during model training. Figure 1(a) illustrates how conditional information $\mathbf{c}$ is converted to the predicted target feature $\tilde{\mathbf{x}}$ by a non-autoregressive decoder. The structure of the decoder is described in [3]. The model is optimized by minimizing the $\mathcal{L}1$ (for spectrogram prediction) or $\mathcal{L}2$ (for prosody prediction) loss between $\tilde{\mathbf{x}}$ and the target feature $\mathbf{x}$. From a probabilistic perspective, minimizing $\mathcal{L}1$ and $\mathcal{L}2$ loss is equivalent to maximizing log-likelihood with Laplacian and Normal distributions. This strong assumption often results in an "over-smoothed" prediction, e.g., flat speech or a blurred image. This loss is typically applied to train both prosody prediction (i.e. duration and $f0$) and spectrogram prediction models.

Normalizing Flows have demonstrated state-of-the-art performance for TTS [9, 10, 16] and voice conversion [16, 17]. For this study, we use the Flow-TTS model topology first described in [9], illustrated in Figure 1(b). Pre-trained phoneme alignments are used instead of using attention following [16–18]. During training, the flow model learns a transformation of the target feature $\mathbf{x}$ into the latent variable $\mathbf{z}$ using a series of invertible flow steps f^{-1}. Conditioning features, $\mathbf{c}$, are provided at each of the flow steps using affine coupling blocks. The flow

Work conducted when all authors were at Amazon TTS Research.

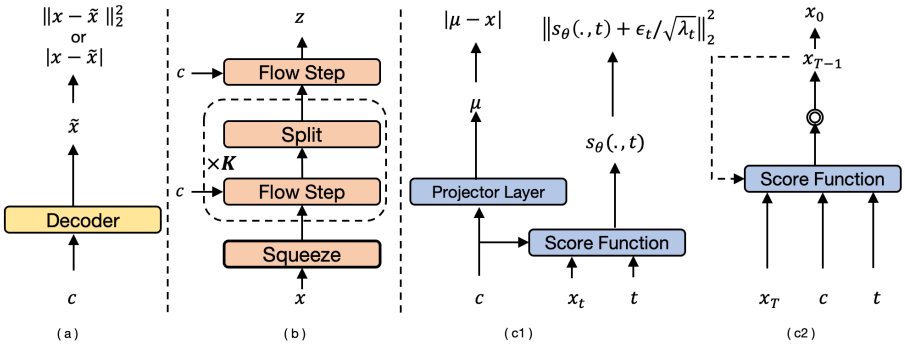

Figure 1: *(a)* ***$\mathcal{L}1$ or $\mathcal{L}2$ Loss-based Model*** *(b)* ***Normalizing Flow Model*** *(c) Training and inference stages of* ***Diffusion Model***

is trained to maximise the likelihood that z comes from a prior distribution. This means the flow learns to map from an unknown complex distribution of x, to a vector which comes from a simpler prior distribution. This allows the model to optimize for the exact log-likelihood of the data distribution. This is in contrast to $\mathcal{L}1/\mathcal{L}2$ losses which place a strong assumption on the distribution of x directly. For this investigation we use a simple $\mathcal{N}(0, 1)$ prior distribution. During inference, the predicted features $\tilde{\mathbf{x}}$ can be derived by sampling from the prior distribution.

Diffusion models [11, 19], similarly to flows, learn to convert from a simple prior distribution to the unknown complex target feature distribution corresponding to the conditional information. Here we omit the mathematical derivation of the diffusion models and refer the reader instead to [11, 19] for further information. During training, noise sampled from the distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I})$ is repeatedly added to the target feature $\boldsymbol{x}$ at each layer of the diffusion model. As a result, eventually x, which comes from an unknown complex distribution, is transformed to the noise distribution as the number of layers $t \rightarrow \infty$. By solving the SDE in [11], the vector $\boldsymbol{x}_t$, output by layer t, can be derived directly without intermediate noisy samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}$.

In parallel, a model is trained to predict the noise that was added at each of the layers of the diffusion network, referred to as the score-based model. The predicted noise is subsequently removed from $\boldsymbol{x}_t$, enabling for a mapping from the noise distribution back to the target data x. During training, the mean of the prior distribution $\boldsymbol{\mu}$ is predicted from the conditional information $\mathbf{c}$ by a projector layer, as shown in Figure 1 (c1). An $\mathcal{L}1$ loss is applied between $\boldsymbol{\mu}$ and x, meaning that the prior of the noise is an over-smoothed spectrogram representation (similar to that learnt by the $\mathcal{L}1/\mathcal{L}2$ approaches described above). Note that we also attempted to use an uninformative $\mathcal{N}(0, 1)$ Gaussian distribution for the noise prior, however the model performed worse and required a larger number diffusion layers. The score-based model $s_\theta(\boldsymbol{x}_t, \mathbf{c}, t)$ is implemented using a U-Net [20] architecture with $\mathbf{c}$, layer t and noisy sample $\boldsymbol{x}_t$ as the inputs. Thus, the loss function consists of two parts, a weighted score-matching objective corresponding to estimating the score function of $p(\boldsymbol{x}_t)$ and an $\mathcal{L}1$ loss between $\boldsymbol{\mu}$ and target feature $\boldsymbol{x}$.

3. Prosody and Acoustic Modelling

3.1. Prosody Models

We investigate the three model architectures described in Section 2 for the task of prosody modelling. We define the task of prosody prediction as the joint modelling of phoneme-level *log-f0* and duration. Frame-level *log-f0* is linearly interpolated over unvoiced regions and then mean-normalized at the speaker level. Given the forced-aligned phoneme sequences, we average *log-f0* at the phoneme-level. Duration is measured by the number of frames aligned to each phoneme.

All models share an identical encoder architecture which processes the input conditioning features. We vary the decoders and optimization steps, following Section 2. The input to the models consists of the phoneme sequence and a categorical speaking style identifier. The encoder for the phoneme sequence follows an identical architecture to the one described in [3]. The one-hot speaking style identifier is transformed by an embedding layer and concatenated to the output of the phoneme encoder. The encoder produces the embedding $\mathbf{c}$, which is used to condition the decoder of the investigated decoder architectures. The prosody models output two-dimensional vectors, corresponding to phoneme-level *log-f0* and duration. We use phoneme-level *log-f0* modelling, as a preliminary analysis found that an acoustic model conditioned with oracle phoneme-level *log-f0* performs slightly better or identically to oracle frame-level *log-f0*. Additionally, the phoneme-level approach has the advantage of allowing joint prediction of *log-f0* and duration, which can better capture the relationship between the two prosodic features.

3.2. Acoustic Models

We investigate the three model architectures in Section 2 also for the task of acoustic modelling. The inputs to all acoustic models are the phoneme sequence, a pre-trained speaker embedding [21], phoneme-level *log-f0* and phoneme durations. Unlike the prosody models, the acoustic models are not conditioned with speaking style information. This is because style is largely conveyed by prosody, also speaker and style attributes are highly entangled in the dataset used. The acoustic models are optimized on the target mel-spectrograms, extracted from the time-domain waveform. As above, the acoustic models share the same encoder, but use differing decoders. The encoder uses the same model architecture as the prosody models. However, unlike the prosody models, the phoneme-level conditional encoding $\mathbf{c}$ is upsampled to the frame level before being passed to the decoder. The speaker embedding and phoneme-level *log-f0* features are concatenated to the phoneme encodings. At training time, oracle phoneme-level *log-f0* and durations are provided to the acoustic model, while at synthesis-time, we use features generated by one of the prosody models.

4. Experimental Protocol

Throughout our experiments, we use a internal dataset of 200 speech hours, recorded by 116 native speakers of English, across a variety of expressive speaking styles such as happiness, sadness, anger, etc. A sampling rate of 24kHz was used for all recordings, from which 80-dimensional mel-spectrograms were extracted with a frame length of 50 ms and a frame shift of 12.5 ms. We use a universal neural vocoder [22] to map generated mel-spectrograms to time-domain waveforms.

To simplify the number of comparisons, we first evaluate the three acoustic model architectures conditioned with oracle prosody features. We then select the best acoustic model and compare the different prosody models. Following Section 3.2, we consider three acoustic models. 1) $\mathcal{L}1$-*AM*: acoustic model trained with $\mathcal{L}1$ Loss. An $\mathcal{L}1$ loss-based model is investigated instead of $\mathcal{L}2$ following recent work [4, 5, 23]. 2) *Flow-AM*: Flow-based acoustic model. 3) *Diff-AM*: Diffusion-based acoustic model. In addition, we define an upper-bound by copy-synthesis, generating time-domain waveforms from oracle mel-spectrograms with the universal neural vocoder. This system is termed *ORA-AM*. Once the best performing acoustic model is selected, we compare the different prosody models. 1) $\mathcal{L}2$-*PM*: prosody model trained with $\mathcal{L}2$ loss. This is selected as it features heavily in recent studies [3, 4, 13, 24]. 2) *Flow-PM*: Flow-based prosody model. 3) *Diff-PM*: Diffusion-based prosody model. As before, an upper-bounded system *ORA-P* is created by feeding the acoustic model with oracle prosody features.

We conduct subjective evaluations of the models using a MUSHRA standard reference evaluation paradigm, considering **Naturalness**, **Style Similarity** and **Expressiveness**. Each listening test included 300 utterances generated by each of the competing systems. Utterances were rated by 300 native speakers using a crowdsourcing platform. Each listener rated 15 MUSHRA screens. We test for statistical significance between systems using paired t-tests with Holm-Bonferroni correction applied. All reported significant differences are at the level of $p < 0.05$. We assign Naturalness and Style Similarity higher importance because we consider higher expressiveness to be favored only if there are no impacts on naturalness and style similarity. We also adopt objective metrics to further analyze the generated prosody features. We observe the standard deviation (STD) of *log-f0*, Δ*log-f0* and duration. These statistics mirror the dynamics of the prosody features that can be associated with the expressiveness of speech [25]. Additionally, we apply the Jensen-Shannon divergence (JSD) [15] to measure the distance between oracle and generated features.

5. Experimental Results

5.1. Acoustic Models

For inference with Flow and Diffusion models, we sample a latent variable from a prior distribution. The temperature τ, i.e., standard deviation, of that distribution can impact the quality of generated speech [10, 11]. Typically, high temperature values, such as $\tau = 1$, bias the model to produce more varied speech, but can negatively impact quality. Meanwhile, a low temperature value often results in flatter intonation [10]. We investigate the temperature which best manages the trade-off of expressivity and quality for the Flow and Diffusion-based models by conducting naturalness subjective evaluations. We consider $\tau \in \{0.2, 0.4, 0.6, 0.8\}$ and present results in Table 1. The highest temperature is chosen when its corresponding Naturalness MUSHRA scores have no statistically significant difference from the highest MUSHRA scores. Therefore, $\tau = 0.4$ and $\tau = 0.8$ were selected for *Flow-AM* and *Diff-AM*, respectively. We also observe that τ has a much larger impact for the Flow-based system than for the Diffusion-based system, implying that careful temperature tuning is especially important for Flow-based systems.

Table 1: *MUSHRA naturalness evaluation results for temperature τ, showing mean values with 95% confidence intervals. * indicates no statistically significant difference from the highest MUSHRA scores.*

System	$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.6$	$\tau = 0.8$
Flow-AM	$\mathbf{77.95 \pm 1.12}$	$77.39 \pm 1.13^*$	77.03 ± 1.14	72.86 ± 1.41
Diff-AM	78.13 ± 1.12	79.20 ± 1.01	$79.79 \pm 1.06^*$	$\mathbf{80.01 \pm 1.05}$

Using the selected temperatures for the Flow- and Diffusion-based systems, we compare all acoustic models conditioned with oracle *f0* and duration. Results for all systems across the three evaluation metrics are presented in Table 2. In terms of naturalness, there are no significant differences between the three acoustic models. *ORA-AM* is significantly preferred over $\mathcal{L}1$-*AM* and *Diff-AM*. However, there is no significant preference between *ORA-AM* and *Flow-AM*. In terms of style similarity, there is no significant difference between *Flow-AM* and $\mathcal{L}1$-*AM*, however both outperform *Diff-AM*. In terms of expressiveness, *Diff-AM* is found to significantly outperform the remaining two systems, while there is no significant difference between *Flow-AM* and $\mathcal{L}1$-*AM*. We hypothesize that *ORA-AM* is rated higher than all three systems in terms of style similarity and expressiveness because, in addition to the oracle prosody features, some speaking styles are expressed by alternative acoustic attributes (e.g., laughing). It is somewhat surprising that *Diff-AM* achieves higher expressiveness but lower style similarity scores than the other two systems. A possible explanation for this is that we are using a higher temperature for the Diffusion-based acoustic model, which may come at the cost of style similarity.

To investigate our results further, we conduct a naturalness preference test on the two best performing systems: $\mathcal{L}1$-*AM* and *Flow-AM*. Relative preference scores for $\mathcal{L}1$-*AM*, *Flow-AM*, and *No Preference* are 24.72%, 29.33%, and 45.95% respectively. A binomial significance test with the *No Preference* scores divided equally amongst the two competing systems indicates a statistically significant preference for *Flow-AM* at the level of $p < 0.05$. Therefore, the Normalizing Flow system is found to provide the best results overall. Consequently, *Flow-AM* is selected to evaluate the prosody models.

Table 2: *Mean MUSHRA scores for acoustic models using oracle f0 and duration, with 95% confidence intervals.*

Method	Naturalness	Style Similarity	Expressiveness
ORA-AM	78.2 ± 1.11	79.75 ± 0.97	83.16 ± 0.83
$\mathcal{L}1$-*AM*	76.87 ± 1.13	77.17 ± 1.07	78.65 ± 0.96
Flow-AM	$\mathbf{77.07 \pm 1.11}$	$\mathbf{77.29 \pm 1.04}$	78.69 ± 0.97
Diff-AM	76.40 ± 1.21	76.20 ± 1.11	$\mathbf{79.49 \pm 0.98}$

5.2. Prosody Models

As before, we begin by finding the best temperature τ for the Flow- and Diffusion-based prosody models. We keep the

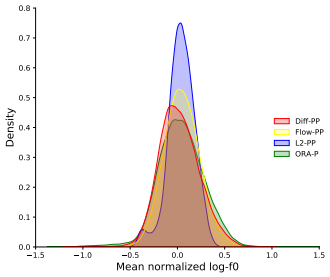

Figure 2: *Distribution of* log-f0 *values generated by the different Prosody Models.*

acoustic model (*Flow-AM*) fixed and condition this with the generated *f0* and duration from the various prosody models. We evaluate the speech samples in terms of naturalness, with results shown in Table 3. Following the same criterion, we choose $\tau = 0.4$ and $\tau = 0.8$ for *Flow-PM* and *Diff-PM*, respectively.

Table 3: *MUSHRA naturalness evaluation results for temperature τ, showing mean values with 95% confidence intervals. * indicates no statistically significant difference from the highest MUSHRA score.*

System	$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.6$	$\tau = 0.8$
Flow-PM	$\mathbf{74.36 \pm 1.10}$	$73.94 \pm 1.12^{*}$	72.28 ± 1.10	70.74 ± 1.10
Diff-PM	$\mathbf{78.40 \pm 0.83}$	$78.27 \pm 0.84^{*}$	$77.92 \pm 0.87^{*}$	$77.71 \pm 0.88^{*}$

We present the results for objective metrics in Table 4. Results find that *Flow-PM* and *Diff-PM* produce features with higher standard deviations than $\mathcal{L}2$-*PM*. This confirms that the proposed prosody models are able to mitigate the over-smoothing problem and produce more dynamic prosody features, which can lead to more expressive speech. Considering JSD, we observe that *Diff-PM* and *Flow-PM* produce features that are closer to the distribution of oracle features. This can also be seen by the *log-f0* distribution from the different prosody models in Figure 2. Most of the *f0* values from $\mathcal{L}2$-*PM* are concentrated around 0. In contrast, the distributions of *f0* from the other two systems are more dispersed and have longer tails, indicating better distribution coverage.

Table 4: *Standard deviation (STD) and Jensen-Shannon divergence (JSD) for the different prosody models.*

System	STD			JSD	
	log-f0	dur	Δ*log-f0*	*log-f0*	dur
ORA-P	0.24	4.49	0.19	-	-
$\mathcal{L}2$-*PM*	0.14	4.03	0.09	0.163	0.112
Flow-PM	0.19	4.12	0.14	0.073	0.067
Diff-PM	**0.21**	**4.27**	**0.15**	**0.057**	**0.041**

Table 5 shows the results from the subjective evaluations of the prosody models. There are no significant differences between the three competing systems and *ORA-P* in terms of naturalness. This finding is perhaps unexpected and suggests that there is little room for naturalness improvement, on average across all speech samples. However, when we consider only utterances from speaking styles with high arousal, such as anger or happiness [26], *Diff-PM* and *Flow-PM* have a larger gap to $\mathcal{L}2$-*PM*, with *Flow-PM* outperforming $\mathcal{L}2$-*PM*. The results indicate that *Flow-PM* and *Diff-PM* can contribute most to prosody modelling for styles with high arousal. In terms of style similarity and expressiveness, no significant differences are found between *Flow-PM*, *Diff-PM* and *ORA-P*. Both *Flow-PM* and *Diff-PM* are found to significantly outperform $\mathcal{L}2$-*PM*. It is somewhat surprising that *Diff-PM* and *Flow-PM* are both on par with *ORA-P* in terms of naturalness, style similarity and expressiveness. Specifically, the objective analysis in Table 4 found the oracle prosody features to have larger standard deviations than those from *Diff-PM* and *Flow-PM*, however it appears as though these differences do not result in listener preferences. A possible explanation for this could be that the expressiveness in speech also depends on how the acoustic model represents the prosody features.

We conducted follow-up preference tests for *Diff-PM* and *Flow-PM* in terms of naturalness, style similarity and expressiveness. However, no significant differences were found between *Flow-PM* and *Diff-PM* for any of the metrics. Overall, *Flow-PM* and *Diff-PM* are on par with each other, but significantly preferred to $\mathcal{L}2$-*PM*.

Table 5: *Mean MUSHRA scores for prosody models, along with 95% confidence intervals.*

Method	Naturalness	Style Similarity	Expressiveness
ORA-P	79.61 ± 0.90	78.62 ± 0.91	77.20 ± 0.92
$\mathcal{L}2$-*PP*	79.02 ± 0.94	76.49 ± 0.98	75.74 ± 0.99
Flow-PM	79.41 ± 0.88	$\mathbf{78.23 \pm 0.91}$	77.17 ± 0.89
Diff-PM	$\mathbf{79.51 \pm 0.92}$	78.10 ± 0.92	$\mathbf{77.32 \pm 0.91}$

6. Discussion

Both flow and diffusion approaches learn a mapping of the target features, coming from complex unknown distributions, transforming them to points from defined simple prior distributions. Losses applied during training are made relative to the likelihood of the prior distribution. However, $\mathcal{L}1$/$\mathcal{L}2$-based models place a strong assumption that the distribution of the target features is a Gaussian and return values that come from the mean of the distribution, resulting in less expressive predictions. We hypothesise that the U-Net structure within the diffusion model is good at generating features considering long-term dependencies, a desirable trait for prosody modelling. However, for acoustic modelling, when provided with prosody conditioning the long-term information is already being largely explained to the model. Instead, the acoustic model is being asked to focus on short-term quality of generated individual spectrogram frames. For such a task it appears that the diffusion model does not perform as well as flow or $\mathcal{L}1$-based models.

7. Conclusion

In this paper, we study and compare three different methodologies for acoustic and prosody modelling: normalizing flows, diffusion probabilistic models, and models trained with $\mathcal{L}1$/$\mathcal{L}2$ loss. For acoustic modelling, subjective evaluation results suggest that an acoustic model based on Normalizing Flows achieves the best results. For prosody modelling, we observe comparable performance for flow-based and a diffusion-based models in terms of naturalness, style similarity and expressiveness. In terms of both objective and subjective evaluation, the prosody features predicted from flow-based and diffusion-based models demonstrate improved expressiveness and better style similarity than the prosody model optimized using an $\mathcal{L}2$ loss.

8. References

[1] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel tacotron: Non-autoregressive and controllable TTS," in *Proc. ICASSP*, 2021, pp. 5709–5713.

[2] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[3] R. Shah, K. Pokora, A. Ezzerg, V. Klimkov, G. Huybrechts, B. Putrycz, D. Korzekwa, and T. Merritt, "Non-autoregressive TTS with explicit duration modelling for low-resource highly expressive speech," *Proc. 11th ISCA Speech Synthesis Workshop (SSW)*, 2021.

[4] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2020.

[5] A. Lańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *Proc. ICASSP*.

[6] G. Zhang, S. Qiu, Y. Qin, and T. Lee, "Estimating mutual information in prosody representation for emotionalprosody transfer in speech synthesis," in *ISCSLP*, 2021.

[7] L. Sheng and E. N. Pavlovskiy, "Reducing over-smoothness in speech synthesis using generative adversarial networks," in *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, 2019.

[8] G. Zhang, Y. Qin, W. Zhang, J. Wu, M. Li, Y. Gai, F. Jiang, and T. Lee, "iEmoTTS: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre," *arXiv preprint arXiv:2206.14866*, 2022.

[9] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-TTS: A non-autoregressive network for text to speech based on flow," in *Proc. ICASSP*, 2020.

[10] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," *Proc. NeurIPS*, vol. 33, pp. 8067–8077, 2020.

[11] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *Proc. ICML*. PMLR, 2021, pp. 8599–8608.

[12] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-TTS: A denoising diffusion model for text-to-speech," *Proc. Interspeech*, 2021.

[13] T. Kenter, V. Wan, C.-A. Chan, R. Clark, and J. Vit, "CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *Proc. ICML*. PMLR, 2019, pp. 3331–3340.

[14] Z. Hodari, O. Watts, and S. King, "Using generative modelling to produce varied intonation for speech synthesis," in *Proc. 10th ISCA Speech Synthesis Workshop (SSW)*, 2019, pp. 239–244.

[15] A. Abbas, T. Merritt, A. Moinet, S. Karlapati, E. Muszynska, S. Slangen, E. Gatti, and T. Drugman, "Expressive, variable, and controllable duration modelling in TTS," *Proc. Interspeech*, 2022.

[16] P. Biliński, T. Merritt, A. Ezzerg, K. Pokora, S. Cygert, K. Yanagisawa, R. Barra-Chicote, and D. Korzekwa, "Creating new voices using normalizing flows," in *Proc. Interspeech*, 2022.

[17] T. Merritt, A. Ezzerg, P. Biliński, M. Proszewska, K. Pokora, R. Barra-Chicote, and D. Korzekwa, "Text-free non-parallel many-to-many voice conversion using normalising flow," in *Proc. ICASSP*, 2022.

[18] A. Ezzerg, T. Merritt, K. Yanagisawa, P. Biliński, M. Proszewska, K. Pokora, R. Korzeniowski, R. Barra-Chicote, and D. Korzekwa, "Remap, warp and attend: Non-parallel many-to-many accent conversion with normalizing flows," in *Proc. SLT*, 2023.

[19] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. ICL*, 2020.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[21] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification," in *Proc. ICASSP*, 2018.

[22] Y. Jiao, A. Gabryś, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, "Universal neural vocoding with parallel wavenet," in *Proc. ICASSP*, 2021.

[23] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "Durian: Duration informed attention network for speech synthesis," 2020.

[24] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Proc. NeurIPS*, vol. 31, 2018.

[25] R. A. Clark, "Using prosodic structure to improve pitch range variation in text to speech synthesis." International Congress of Phonetic Sciences, 1999.

[26] E. A. Kensinger and D. L. Schacter, "Processing emotional pictures and words: Effects of valence and arousal," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 6, no. 2, pp. 110–126, 2006.