Non-Autoregressive TTS with Explicit Duration Modelling for Low-Resource Highly Expressive Speech

Raahil Shah[*]*, Kamil Pokora*[*]*, Abdelhamid Ezzerg, Viacheslav Klimkov, Goeric Huybrechts, Bartosz Putrycz, Daniel Korzekwa, Thomas Merritt*

Amazon Text-to-Speech Research

{raahshah, kamipoko}@amazon.com

Abstract

Whilst recent neural text-to-speech (TTS) approaches produce high-quality speech, they typically require a large amount of recordings from the target speaker. In previous work [1], a 3-step method was proposed to generate high-quality TTS while greatly reducing the amount of data required for training. However, we have observed a ceiling effect in the level of naturalness achievable for highly expressive voices when using this approach. In this paper, we present a method for building highly expressive TTS voices with as little as 15 minutes of speech data from the target speaker. Compared to the current state-of-the-art approach, our proposed improvements close the gap to recordings by 23.3% for naturalness of speech and by 16.3% for speaker similarity. Further, we match the naturalness and speaker similarity of a Tacotron2-based full-data ($\approx$ 10 hours) model using only 15 minutes of target speaker data, whereas with 30 minutes or more, we significantly outperform it. The following improvements are proposed: 1) changing from an autoregressive, attention-based TTS model to a non-autoregressive model replacing attention with an external duration model and 2) an additional Conditional Generative Adversarial Network (cGAN) based fine-tuning step.

Index Terms: Text-to-speech, low-resource, expressive speech

1. Introduction

Recent advancements in the TTS domain have demonstrated highly natural speech generated by neural text-to-speech (NTTS) models [2, 3, 4, 5, 6]. However, these models often require large amounts ($\approx$ 10 hours) of recordings [7] to achieve high levels of naturalness without degradation.

Data collection for TTS is an expensive and time-consuming task. The problem is magnified for highly expressive voices, because it requires higher vocal effort from the voice talent as compared to neutral speech. This amplifies the need for a scalable solution to be able to build highly expressive voices with smaller amounts of data and without substantial cost (i.e. low-resource TTS).

Previous research around low-resource TTS attempts to address this problem with multi-speaker modelling and transfer learning. Transferring knowledge from full-resource speakers to a low-resource one improves the synthesis quality of the low-resource speaker [7, 8, 9, 10, 11, 12, 13].

Recent work in Huybrechts et al. [1] brings significant improvements to naturalness by combining multi-speaker modelling with data augmentation for the low-resource speaker. This approach uses a Voice Conversion (VC) model [14, 15, 16, 17, 18] to transform speech from one speaker to sound like speech from another, while preserving the content and prosody of the source speaker. This artificially boosts the training data available for the resource-scarce target speaker by leveraging readily available source speaker data. However, we have observed that this solution does not scale to achieve naturalness on par with a full-data model for more expressive voices than those presented in [1].

To address this limitation, we investigate the most expressive voice in our catalog and propose changes to the model architecture that consistently outperform the approach presented in [1] and achieve naturalness on par or better than a full-data Tacotron2-based [2] model.

First, we propose to switch from a Tacotron2-based (autoregressive) TTS model to a non-autoregressive mel-spectrogram prediction model and to replace the attention mechanism in Tacotron2 with an external duration model. To the authors' knowledge, this work is the first to investigate such NTTS architectures in a reduced data scenario. In the literature, so far mainly attention-based or autoregressive models have been explored in the context of expressive low-resource TTS [8, 19, 20, 9]. Such models suffer from stability issues exhibited in synthesised speech, such as babbling, early cut-off, word repetition, and word skipping [21, 22, 23, 24]. These problems, attributed to teacher-forcing and attention, are even more prevalent in the reduced data scenario. Recent research in the field [25, 26, 27, 28], inspired by traditional parametric speech synthesis [29, 30] mitigates these issues by explicitly modelling the durations of phonemes. In addition to improving speech stability, we posit that explicit duration modelling significantly improves the overall naturalness of highly expressive voices by making it easier to model variability in phoneme durations than in the baseline attention-based systems.

Second, we investigate an application of Conditional Generative Adversarial Networks (cGAN) [31] as an additional fine-tuning step aimed at improving the signal quality of low-resource synthesis. The less data we have, the harder it is to maintain good segmental quality and speaker similarity. GANs [32], known for generating high quality images, have also been applied in the speech domain to improve the segmental quality of predicted mel-spectrograms [33, 34]. We extend the standard GAN recipe to pass conditioning in addition to the typical mel-spectrogram input to the discriminator. This better informs the discriminator network when making a classification, allowing for more insightful information to flow to the generator.

2. Proposed Method

As in Huybrechts et al. [1], the method presented in this paper is based on three main steps: 1) data augmentation, 2) multi-speaker TTS and 3) fine-tuning. In this work, we also investigate the addition of a fourth step where we fine-tune the model with a cGAN approach to further improve the audio quality.

[*]The first two authors have equal contribution.

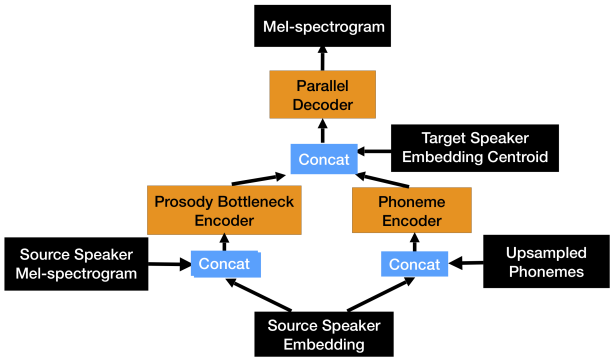

Figure 1: *Schematic diagram of the voice conversion model used in Step 1 of the proposed method.*

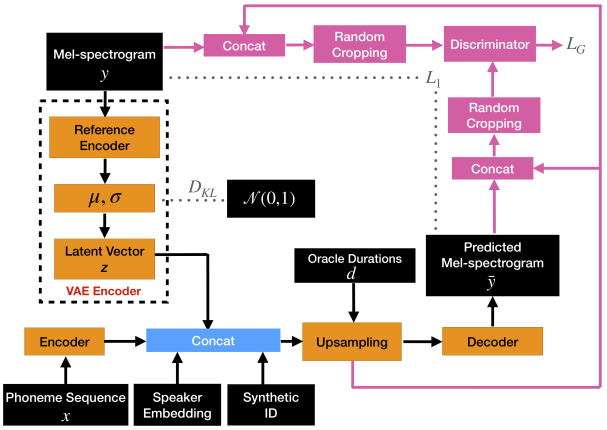

Figure 2: *Schematic diagram of the acoustic model used in Steps 2-4 of the proposed method. Components in pink are used only in Step 4 of the method.*

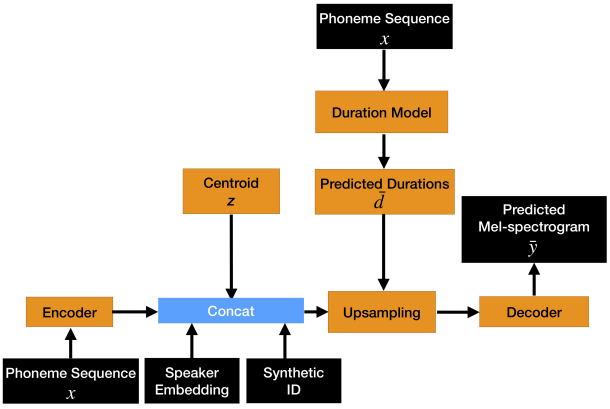

Figure 3: *Schematic diagram of the acoustic model during inference.*

Our full proposed low-resource TTS methodology is defined as follows:

1. Train a VC model to augment data for the target speaker.
2. Train a multi-speaker TTS model using recordings and synthetic data created in Step 1.
3. Fine-tune the TTS model with the recordings from the target speaker.
4. Fine-tune the TTS model with the cGAN approach.

The key contribution of this work is the change in TTS architecture from a Tacotron2-style attention-based model to a non-autoregressive acoustic model supported by external durations. The resulting TTS model is comprised of two main components: 1) an acoustic model that predicts mel-spectrogram $\tilde{y}$ from a phoneme sequence x, 2) a duration model which assists the acoustic model during inference by providing the duration $\tilde{d}$ of each phoneme. During training, ground truth durations d are used by the acoustic model. As in Huybrechts et al. [1], a Parallel WaveNet universal neural vocoder [35] is used to obtain the final speech signal from the generated mel-spectrogram.

2.1. Voice Conversion Model

A voice conversion model is used to perform data augmentation in Step 1 of the method. This model converts the speaker identity of a source audio to sound as though it was spoken by the target speaker.

As in Huybrechts et al. [1], we use the CopyCat [18] architecture for this model which is presented in Figure 1. The model consists of: 1) a phoneme encoder that learns latent representations from phonemes, 2) a prosody bottleneck encoder which disentangles prosody from the reference mel-spectrogram and 3) a parallel decoder which generates the mel-spectrogram given the phoneme and prosody bottleneck encoder's outputs, in addition to the target speaker embedding.

We follow the approach in Huybrechts et al. [1] to modify the original CopyCat model by concatenating speaker embeddings to the upsampled phonemes before feeding this to the phoneme encoder. This was found to help reduce occurrences of speaker leakage in [1].

The VC model was trained with 18 supporting speakers who were recorded in a conversational speaking style, in addition to the target speaker. For the highly expressive target speaker investigated in this paper, fine-tuning of the Copycat model was required to prevent issues with speaker leakage, unlike in Huybrechts et al. [1]. The model was trained on the data from all speakers for 50k steps and then only on the target speaker's data for an additional 320 epochs. We hypothesise that this fine-tuning is required because the target speaker's data is much more expressive than that of the supporting speakers.

2.2. Acoustic Model

We use an acoustic model in Step 2-4 of the method, as illustrated in Figure 2. The topology of this model during inference is presented in Figure 3.

2.2.1. Encoder

Our encoder architecture is the same as that presented in Tacotron2 [2]. It is comprised of an embedding lookup followed by 3 convolution blocks each with a kernel size of 3. On top of that we apply a single bi-directional LSTM layer with a hidden dimension of 512 and a dropout of 0.1. We pass the phoneme sequence x as input to this encoder, to obtain phoneme embeddings $\tilde{x}$.

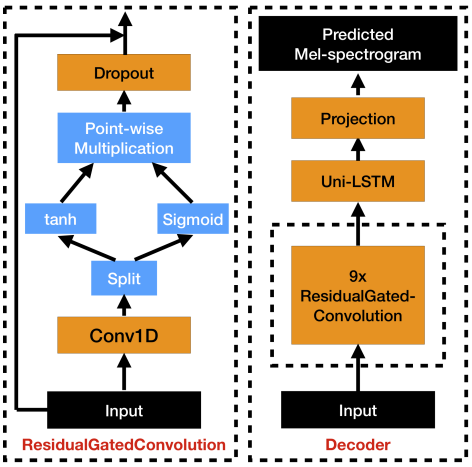

Figure 4: *Illustration of a residual gated convolution block and the decoder architecture.*

2.2.2. *Variational Autoencoder (VAE)*

TTS is a one-to-many problem as the same text can be spoken in many different, yet acceptable, ways. In autoregressive NTTS models, this effect is mitigated both by teacher-forcing as well as by conditioning on the latent acoustic representation obtained from a VAE [36]. In the proposed non-autoregressive architecture, we use only the VAE to pass information which cannot be inferred solely from the input phoneme sequence.

This encoder takes mel-spectrogram frames as input. It comprises of 6 convolution blocks each with a kernel size of 5, followed by one GRU layer with a hidden dimension of 128. We take the last output from the GRU and perform a projection to 128 dimensions, in order to parametrise the posterior distribution. The first half of this output represents μ while the second half represents σ. Finally, we sample from the posterior distribution to obtain a final latent representation z. At inference time, we use a pre-calculated centroid of zs of the available ground truth data for the target speaker.

2.2.3. *Upsampling and Additional Embeddings*

To each phoneme embedding we concatenate: 1) the latent z vector, 2) a speaker embedding obtained from a pre-trained GE2E-based [37] speaker verification model and 3) a one-hot 'synthetic ID' flag, indicating whether the data is ground truth or obtained from voice conversion. Then we upsample each phoneme embedding according to ground truth durations d (training-time) or predicted durations $\tilde{d}$ (inference-time).

Similar to Parallel Tacotron [27], before passing these upsampled embeddings to the decoder, we provide positional information to indicate the relative position of a frame inside a phoneme. To each embedding we concatenate: 1) a transformer-style positional embedding [38] indicating the phoneme duration, 2) a transformer-style positional embedding indicating the frame's position inside a phoneme and 3) the fractional progress of the frame in a phoneme.

2.2.4. *Decoder*

The embedding sequence output from the upsampling component is passed as input to the decoder. The modelling task of the decoder was found to require local context in [27], therefore our decoder is comprised of 9 residual gated convolution layers. Each residual gated convolution block is composed of a 1D-convolution with kernel size 15 and a hidden dimension of 512, followed by a $\tanh$ filter and sigmoid activation gate which are element-wise multiplied and then added to a residual connection after a dropout of 0.1.

The convolution stack is followed by 2 uni-directional LSTM layers with a hidden dimension of 512 and a dropout of 0.1. Preliminary evaluations showed that this final LSTM stack improves audio quality. A schema of the decoder as well as the residual gated convolution architecture is presented in Figure 4.

2.2.5. *Conditional GAN Fine-Tuning*

GANs are a well established solution to the problem of 'over-smoothing' encountered during the optimisation of L1/L2 loss functions. With mel-spectrogram prediction, this effect manifests as lower brightness and poorer audio quality in the subjective perception of the speech signal.

Adversarial training of the acoustic model can be utilised as a fine-tuning step to mitigate such degradations [33, 34]. Typically, such an adversarial training involves only the mel-spectrogram being passed as input to the discriminator network. We explore an extension to this setup (cGAN), wherein we condition the discriminator on both acoustic and linguistic information. Additional conditioning allows for more meaningful gradient flow from discriminator to generator, which has been shown to improve adversarial training [31].

The entire acoustic model acts as the generator network. For the discriminator network we used the architecture presented in SAGAN [39]. As input to the discriminator we feed randomly cropped 64 frame chunks of the generated mel-spectrogram, along with the embeddings $\tilde{x}$ of the corresponding phoneme sequence and the latent acoustic information z from the VAE. Cropping was found to be more effective than feeding the whole mel-spectrogram. We hypothesise that this is because the goal of the fine-tuning step is to improve the segmental quality of the final mel-spectrogram, which is a more local, time-invariant task.

2.2.6. *Training Setup*

To train the acoustic model we use the Adam optimiser [40] with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We use a linear warm-up of the learning rate from 0.1 to 1 for the first 10K steps, followed by an exponential decay from 10K steps to 100K steps with a minimum value of 10^{-5}.

In Step 2 of the method, the acoustic model is trained for 500K steps with a mini-batch size of 32. The model is trained on both ground truth and synthetic data for our target speaker as well as data from supporting speakers, using the following loss function:

$$L_{Train} = L_1 + \gamma * D_{KL} \quad (1)$$

where L_1 is the L_1-distance between predicted and ground truth mel-spectrogram and D_{KL} is the Kullback–Leibler divergence between the VAE posterior distribution and $\mathcal{N}(0, 1)$. To avoid the collapse of D_{KL}, we used the same KL annealing scheme as presented in [41].

In Step 3 of the method, we fine-tune the model for an additional 30K training steps, using only ground truth data from the target speaker, still optimising L_{Train}.

Finally, in Step 4 of the method, we freeze all VAE weights and fine-tune the acoustic model with the cGAN setup for an additional 30K steps, also using only ground truth target speaker

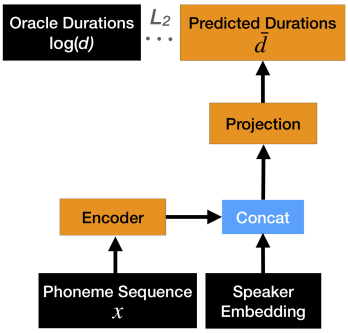

Figure 5: *Schematic diagram of the duration model used in Step 2 of the proposed method.*

data. During this step, the following generator loss (L_G) and Hinge discriminator loss (L_D) functions are used:

$$L_G = \mathbb{E}_{x,y\sim p_{data}}[D(y,\tilde{x},V(y)) - D(G(x,y),\tilde{x},V(y))] \quad (2)$$

$$L_D = \mathbb{E}_{x,y\sim p_{data}}[\text{ReLU}(1 + D(G(x,y),\tilde{x},V(y))) + \text{ReLU}(1 - D(y,\tilde{x},V(y)))] \quad (3)$$

where D is the discriminator network, G is the generator network (acoustic model) and V denotes the VAE. The discriminator is trained by optimising L_D, while the acoustic model is fine-tuned by optimising the total loss:

$$L_{GANFineTune} = L_1 + \alpha * L_G \quad (4)$$

2.3. Duration Model

We train a duration model in Step 2 of the method, whose architecture is presented in Figure 5. We model phoneme durations as the integer number of mel-spectrogram frames corresponding to each phoneme. We assume that ground truth phoneme durations are provided by an external aligner, such as the Gaussian Mixture Model (GMM) based Kaldi Speech Recognition Toolkit [42] used in our experiments.

To model the duration sequence, we first pass the phoneme sequence through an encoder and then apply a dense projection to 1 dimension followed by a ReLU activation function. During training, teacher-forcing is used i.e. only ground truth durations are input to the acoustic model, while predicted durations are used only at inference-time.

For the proposed multi-speaker acoustic model using reduced target speaker data, we train the duration model separate from the acoustic model. This model uses a phoneme encoder identical to the one described in Section 2.2.1, with the hidden dimension reduced to 256, whose output is concatenated with pre-trained speaker embeddings after they are passed through an affine layer. The training objective for this model is a L2 loss on durations in the log domain and it is trained for 150K steps (with a mini-batch size of 32). We use an identical Adam optimiser configuration as that used for the acoustic model training.

For the full-data anchor models trained on a single-speaker, the embedding sequence used for duration prediction is the concatenation of the phoneme embeddings $\tilde{x}$ and the latent vector z from the acoustic model. In this setup, in line with the state-of-the-art [26, 27, 28], the two models are trained jointly, by adding an auxiliary L1 loss between ground truth and predicted durations to the total objective, with a weighting of 0.025.

Preliminary evaluations showed that separately training the duration model in the low-resource multi-speaker scenario performed better than the joint training of acoustic and duration models used for full-data single-speaker models.

3. Experiments

3.1. Data

For the 'highly expressive' target speaker, we selected the voice objectively identified as the most expressive speaker in our internal American English voice catalog. Expressivity was measured as variation along the three axes of frequency, power and durations by analysing respectively the mean and variance of static $\log f0$, $mgc0$ and phoneme duration features and their deltas, for each speaker.

In the 'full-data' (FD) setup we used $\approx$ 10 hours of recorded speech from the target speaker. In the low-resource aka 'data reduction' (DR) scenario, we investigated four different reduced data amounts: 3 hours, 1 hour, 30 minutes and 15 minutes of target speech. To perform data augmentation as detailed in Section 2.1, we supplemented the target speaker data in each DR scenario with 4.5 hours of synthetic data converted from a single source speaker, by a VC model trained using the respective reduced data amount for that scenario.

For supporting speakers in our multi-speaker models, we used an internal American English dataset comprising of 18 speakers recorded in a conversational style. This dataset contained $\approx$ 65 hours of speech.

3.2. Evaluation

In each DR scenario, we evaluated our models by conducting MUSHRA tests [43] on the following two metrics:

- Naturalness – "Please rate the audio samples in terms of their naturalness".
- Speaker Similarity – "Please listen to the speaker in the reference sample first. Then rate how similar the speakers in each system sound compared to the reference speaker."

Each test was conducted independently, by 20 listeners, each evaluating 61 MUSHRA screens synthesised from a fixed test set of 61 held-out samples. To check for statistical significance, we performed paired t-tests using the Holm-Bonferroni correction method. All statistical differences presented are for $p \leq 0.05$.

3.3. GAN Fine-Tuning Study

We conducted a supporting study to investigate improvements from GAN fine-tuning on the naturalness of synthesised speech, demonstrating the impact of Step 4 of the proposed method. In this study, we evaluated the following four systems in a 3 hours DR scenario:

- *(Recordings)* Ground truth recordings.
- *(DR No-att)* Baseline without any GAN fine-tuning (i.e. Steps 1-3 of the proposed method).
- *(DR No-att + GAN)* Candidate system with additional fine-tuning using vanilla GAN (i.e. only mel-spectrogram input to discriminator).
- *(DR No-att + cGAN)* Candidate system with additional fine-tuning using Conditional GAN (i.e. conditioned on the phoneme sequence, acoustic and prosody information).

Target Data	3 h
Naturalness	
Recordings	88.94
DR No-att	64.45
DR No-att + GAN	64.08
DR No-att +cGAN	66.57

Table 1: *Average MUSHRA scores for naturalness, showing the impact of Conditional GAN fine-tuning.*

As shown in Table 1, cGAN fine-tuning provides a statistically significant improvement to naturalness when compared to both vanilla GAN fine-tuning and the baseline without GANs. This demonstrates that the addition of conditioning information does indeed appear to help the discriminator make better distinctions of whether a sample is real or fake, which in turn leads to improvements in the samples produced by the generator.

3.4. Data Reduction Study

Our primary study investigates the impact of the proposed changes to the model architecture and methodology presented in Huybrechts et al. [1], on different amounts of reduced data for the highly expressive target speaker.

In this study, we evaluated the following candidate DR systems: *'DR No-att + cGAN'* and *'DR No-att'*, i.e. the proposed non-autoregressive, external duration TTS model with and without cGAN fine-tuning respectively. We compared them against a baseline DR system to investigate the ablation of our proposed architectural changes and against full-data anchor systems to investigate the ablation of data amount.

'DR baseline' denotes the system presented in Huybrechts et al. [1] which has been shown to synthesise high quality, expressive voices from as little as 15 minutes of data. The same synthetic data is used in the training of both candidate and baseline DR systems.

As full-data anchor systems we used: 1) *'FD Tacotron2'* – a Tacotron2-based TTS model and 2) *'FD No-att'* – the proposed non-autoregressive TTS model. Both full-data systems used an utterance-level VAE and were single-speaker, i.e. trained on all 10 hours of data from the target speaker.

Target Data	3 h	1 h	30 min	15 min
Naturalness				
Recordings	85.70	83.61	86.39	82.30
FD No-att	64.17	65.01	61.41	69.27
FD Tacotron2	58.35	58.88	55.07	63.37
DR No-att	62.80	64.96	59.16	64.31
DR No-att + cGAN	64.94	65.29	59.33	64.22
DR baseline	54.40	59.86	51.21	58.73
Speaker similarity				
Recordings	91.94	92.47	94.04	95.95
FD No-att	69.90	65.21	64.85	70.37
FD Tacotron2	64.11	59.90	58.66	64.05
DR No-att	69.14	66.75	62.94	65.97
DR No-att + cGAN	71.54	66.72	62.95	66.21
DR baseline	63.71	60.74	53.58	60.43

Table 2: *Average MUSHRA scores for naturalness and speaker similarity, showing the performance of proposed method in the context of different amount of data. Note that each column is made up of MUSHRA evaluations for one particular data amount, thus scores are not comparable across different columns. Underlined values signify the best performing system amongst DR systems, up to statistically significant differences.*

The results of this study are presented in Table 2. They show that in terms of naturalness and speaker similarity, the proposed method, *'DR No-att + cGAN'* significantly outperforms the state-of-the-art approach from Huybrechts et al. [1] (i.e. *'DR Baseline'*) for every data amount, demonstrating a clear improvement to low-resource TTS. Improvements from the proposed changes to the model architecture are further highlighted by the result that *'DR No-att + cGAN'* significantly outperforms *'FD Tacotron2'* when there is 30 minutes or more of target speaker data (up to 95% data reduction) and matches it in the 15 minutes scenario.

Compared to *'FD No-att'*, a full-data model with similar architecture, *'DR No-att + cGAN'* is on par for naturalness while bringing a significant improvement to speaker similarity in the 3 hour scenario and is on par for both metrics in the 1 hour scenario. These results highlight the strength of the proposed 4 step methodology in compensating for the reduction in training data.

The gap between *'DR No-att'* and *'DR No-att + cGAN'* diminishes as we reduce the data further, suggesting that the audio quality improvements brought about from Step 4 (cGAN fine-tuning) are statistically significant only when a relatively large amount of data (3 hours) is available for the target speaker.

4. Conclusions

We proposed improvements to the state-of-the-art low-resource TTS technology presented in Huybrechts et al. [1], addressing its limitations when applied to highly expressive voices. The improvements were to: 1) model architecture, i.e. the switch to a non-autoregressive acoustic model supported by external durations in favour of an attention-based, autoregressive Tacotron2 architecture, and 2) methodology, i.e. an additional cGAN fine-tuning step.

The proposed system significantly outperforms the state-of-the-art in both naturalness and speaker similarity, closing the gap to recordings by 23.3% and 16.3% respectively, using as little of 15 minutes of speech from the target speaker. Further, compared to a Tacotron2-based model trained on full-data ($\approx$ 10 hours of speech), the proposed model is on par with just 15 minutes of target speaker data and significantly improves naturalness and speaker similarity with 30 minutes or more data. Finally, with 3 hours of target speaker data, our proposed architecture with additional cGAN fine-tuning outperforms even a full-data model of a similar architecture.

These contributions demonstrate a robust NTTS method that can build high quality, natural speech from as little as 15 minutes of target speaker data and can scale even to highly expressive voices. Such a method can save substantial cost and time invested in data collection for TTS.

Future work includes applying the proposed method to more voices that are challenging to model, such as expressive multi-lingual or character voices. Further, we intend to explore fine-grained prosody embeddings to better model and control expressive speech. We also intend to investigate the joint training of acoustic and duration models for the multi-speaker DR scenario, which was found to underperform compared to the separate training approach presented in this paper.

5. References

[1] G. Huybrechts, T. Merritt, G. Comini, B. Perz, R. Shah, and J. Lorenzo-Trueba, "Low-resource expressive text-to-speech using data augmentation," 2020.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *arXiv preprint arXiv:1712.05884*, 2017.

[3] J. Sotelo *et al.*, "Char2wav: End-to-end speech synthesis," 2017.

[4] R. Skerry-Ryan *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *arXiv preprint arXiv:1803.09047*, 2018.

[5] N. Kalchbrenner *et al.*, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[6] A. Oord *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 3918–3926.

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

[8] A. Gibiansky *et al.*, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in neural information processing systems*, 2017, pp. 2962–2970.

[9] Y. Jia *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, 2018, pp. 4480–4490.

[10] N. Tits *et al.*, "Exploring transfer learning for low resource emotional tts," in *Proceedings of SAI Intelligent Systems Conference*. Springer, 2019, pp. 52–60.

[11] J. Latorre *et al.*, "Effect of data reduction on sequence-to-sequence neural tts," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7075–7079.

[12] Y.-J. Chen *et al.*, "End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning." in *Interspeech*, 2019, pp. 2075–2079.

[13] H. Zhang *et al.*, "Unsupervised learning for sequence-to-sequence text-to-speech for low-resource languages," *arXiv preprint arXiv:2008.04549*, 2020.

[14] S. H. Mohammadi *et al.*, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[15] C.-C. Hsu *et al.*, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[16] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *arXiv preprint arXiv:1804.04262*, 2018.

[17] T. Kaneko *et al.*, "Cyclegan-vc2: Improved cyclegan-based nonparallel voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6820–6824.

[18] S. Karlapati *et al.*, "Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech," *arXiv preprint arXiv:2004.14617*, 2020.

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

[20] Z. Zhang, Q. Tian, H. Lu, L.-H. Chen, and S. Liu, "Adadurian: Few-shot adaptation for neural text-to-speech with durian," *arXiv preprint arXiv:2005.05642*, 2020.

[21] M. He, Y. Deng, and L. He, "Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS," 2019, pp. 1293–1297.

[22] Y. Zheng, J. Tao, Z. Wen, and J. Yi, "Forward–backward decoding sequence for regularizing end-to-end tts," *IEEE/ACM Trans. Audio Speech & Lang. Process.*, vol. 27, no. 12, pp. 2067–2079, 2019.

[23] H. Guo, F. K. Soong, L. He, and L. Xie, "A New GAN-Based End-to-End TTS Training Algorithm," in *Proc. Interspeech*, 2019, pp. 1288–1292.

[24] E. Battenberg, R. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-relative attention mechanisms for robust long-form speech synthesis," in *Proc. ICASSP*, 2020, pp. 6194–6198.

[25] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech," *arXiv:1905.09263*, 2019.

[26] 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," *arXiv:2006.04558*, 2020.

[27] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. Weiss, and Y. Wu, "Parallel tacotron: Non-autoregressive and controllable tts," *arXiv preprint arXiv:2010.11439*, 2020.

[28] J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-Attentive Tacotron: Robust and Controllable Neural TTS Synthesis Including Unsupervised Duration Modeling," *arXiv:2010.04301*, 2020.

[29] H. Zen, K. Tokuda, and A. Black, "Statistical Parametric Speech Synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[30] H. Zen, A. Senior, and M. Schuster, "Statistical Parametric Speech Synthesis Using Deep Neural Networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[31] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[32] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprints*, 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[33] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 4910–4914.

[34] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative adversarial network-based postfilter for stft spectrograms," in *The Annual Conference of the International Speech Communication Association (Interspeech)*, August 2017.

[35] Y. Jiao, A. Gabrys, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, "Universal neural vocoding with parallel wavenet," 2021.

[36] D. P. Kingma *et al.*, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[37] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[39] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018.

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

[41] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," 2016.

[42] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[43] R. ITU-R, "1534-1, method for the subjective assessment of intermediate quality levels of coding systems (mushra),"," *International Telecommunication Union*, 2003.