Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis Reproducibility Challenge

Dan Dascalescu djd2g18@soton.ac.uk
Chalothorn Chavalitcheevinkul cc1f20@soton.ac.uk
Doga Keskin dk1n21@soton.ac.uk
Pasit Kittawarnrat pk2n21@soton.ac.uk

***Abstract*—Flowtron is an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis, an adaptation on existing autoregressive approaches using flows.**

Flowtron optimises training data as stable and straightforward by maximising the likelihood of the training data. The claimed Mean Opinion Scores (MOSs) indicate an improvement over standard autoregressive approaches, both in regular speach and style-transfer[1]. In this work, we train similar models to assess the reproducability of the results in Flowtron, comparing speech outputs to those of Tacotron2 and human speech as a control.

Reproducibility challenge Flowtron repo

I. Introduction

Speech synthesis models such as Flowtron are generative models that learn probabilistic mappings from text to Mel-Spectrograms after training on large batches of labelled speech audio. As they are autoregressive, they generate samples based on prior information (RNNs) so probabilities for the next sample consider all previous samples. In the output (after converting Mel-spectrogram to speech), this manifests in speech samples that vary based on the audio (text) that was generated before — so a word may be said differently if it follows different words, which is a result that can be exploited to produce variations in the styles of speech that are generated, considering the training data.

The mechanism of Flowtron minimises the exact likelihood of the training data to make the training data stable and straightforward. It learns an invertible mapping from data to latent space via normalising flows, which may be adjusted to influence many features of voice synthesis.

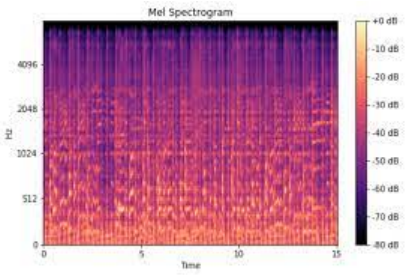

Fig. 1. Mel-Spectrograms: temporal representations of speech

II. Definitions

Before evaluating the reproducibility experiments, we cover the models and methods used.

A. WaveGlow

WaveGlow is a separate (and separately trained) model that produces sound by sampling Mel-Spectrograms [2]. This, when used in conjunction with Flowtron, or Tacotro, creates an end-to-end text-to-speech system as desired. These models are trained simply on unlabeled speech data, and generally transfer very well between voice types, accents, and even languages. Therefore, a custom WaveGlow model is not always required as opposed to simply using pretrained models. WaveGlow is a flow-based neural network that synthesis an audio file based on Gaussian distribution by using mel-spectrogram conditioning. In each process of WaveGlow, there are invertible convolution combining with WaveNet architecture.

B. Tacotron 2

Tacotron is an older (2018) and similar architecture that Flowtron was based on [6]. The method proposed in Tacotron 2 paper is a recurrent sequence-to-sequence feature prediction network that maps character embeddings to mel-spectrograms. A modified WaveNet model works as a vocoder to generate time-domain waveforms from spectrograms. To ensure Tacotron 2 design choices, provide ablation tests of essential system components and assess the impact of employing mel-spectrograms as WaveNet conditioning input instead of language, duration, and F0 attributes. Tacotron 2 demonstrates that utilising tiny acoustic intermediate representation reduces the size of the WaveNet architecture significantly.

C. Flowtron

We implement our experiment using Flowtron. Use latent variable sampling from a known distribution normalizing flow to generate samples and produce a sample for the target distribution by applying invertible transformations. The latent distribution used in Flowtron is a zero-mean spherical Gaussian and a combination of spherical Gaussian with trainable parameters. The invertible function is one of the vital parts that is causal by the autoregressive affine coupling layer.

The text encoder in the Flowtron model architecture is modified from the text encoder in Tacotron 2 to replace batch-norm with instance-norm. The decoder in Flowtron removes the Prenet and Postnet layers from Tacotron[3] to be essential instead.

D. Mean opinion scores (MOSs)

MOSs are subjective scores to assess quality of voice and video, they are widely used as it is the most effective method of evaluating models like these (them being intended for human use, after all) [4]. MOS is the average of a variety of other human-scored individual parameters and is usually graded on a scale of 1(bad) to 5(excellent). Objective media quality metrics rely on data from these subjective experiments for tuning and validation and are, therefore, affected by the same choices and factors. However, the scores of MOS are prone to misuse or misinterpretation. Choices made during subjective media quality test design have a significant impact on MOS values and must be considered when analysing and interpreting MOS data.

III. Experiments

The purpose of this overall experiment is to assess the reproducibility of the results presented in [1]. To that end, we train a custom model, as closely in definition to the paper as possible, using similar datasets and configurations, and (once inference is performed) produce MOS scores based on the audio using subjective testing using volunteers. We compare these to those from Tacotron2 (obtained similarly) and to the ground truth (human speech). The comparisons made between models will then themselves be compared to those made in [1].

We train a Flowtron model, and Tacotron2 model for comparison using the 'Jane Eyre' subset of the en_UK subset of the M-AILABS Speech Dataset [7]. Both of these models take a very long time to train and the original authors of the paper trained them using 8 specialised GPUs, to shorten the training time we start the training from a Flowtron model pre-trained on the LJSpeech dataset for faster convergence. We also performed fine-tuning of models after additional training using a LibriTTS subset, which allowed for better transfer between voices as the LJS dataset (that the majority of training was done with) is a single female US speaker. Nonetheless, it can be expected (and has been observed) that transfer learning from these datasets often performs better when the target voice is female, which was one reason we (as well as the authors of [1]) elected for a transfer set of a female voice.

The authors of [1] included in their training a proprietary dataset consisting of two speakers with a duration of 30 combined hours, this allowed them to have two prominent voices in their experiments and they could interpolate between speaker styles. To replicate the prominent voice effect we used a single speaker, however we do not have a second prominent speaker to easily interpolate between. We chose data spanning both the UK and US accents to assess whether either would be more successful.

Sample audio files and many configuration files and custom training scripts can be found here.

A. Datasets

LibriTTS: A large 585 hours long dataset of English speech with many speakers, with annotations. It is used in one of the pre-trained Flowtron models.

M-AILABS Speech Dataset: A multi-lingual annotated speech dataset, we use a small subset of this dataset to train our own models using fine-tuning. Note that the sampling rate is 16kHz (during the experiments we upsampled this to match the 22050Hz recommendation from Flowtron).

LJ Speech: A dataset consisting of annotated short audio clips from a single speaker with a total length of 24 hours. This dataset is used in the original Flowtron paper, and creates the main pre-trained Flowtron model.

B. Flowtron training

Flowtron was trained for between 50-10000 iterations (2-400 epochs using a batch size of 2 and 50 training samples) using 4x GTX1080ti GPUs. Parameters were optimized with the Adam optimizer (RAdam). A learning rate of 1e-4 and weight decay of 1e-6 were used, as per [1]. Sigma value used during inference was 0.5-0.8, depending on file.

As mentioned, we elected to fine-tune models pre-trained on the LJS and LibriTTS datasets. As such, we fine-tune the learned weights in the speaker embeddings layer. Due to the nature of speech, training procedure is to observe the loss is converging (occasionally, if configurations are incorrect, the training will diverge wildly and produce either white noise or nothing) and to train the model until the attention figures look good (see Fig. 3). Another facet is that training for too long on a small speaker set will overfit dramatically — the attention dissipating (see Fig. 4) and the outputs producing nonsense speech.

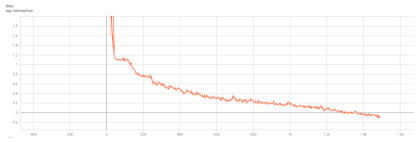

Fig. 2. Expected loss plot: notably converging on -1, not 0

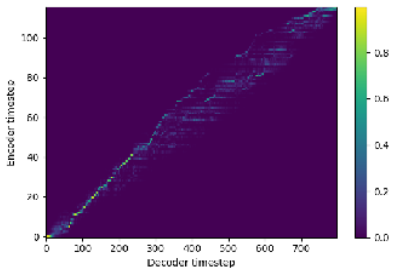

Fig. 3. Expected attention, though not fully converged

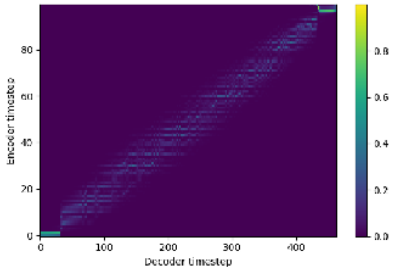

Fig. 4. An attention graph showing model divergence

In the case of Flowtron, we had difficulty fine-tuning to the UK accent and had better luck with the US accented training sets. We found that the model converged better when it had experience of the LibriTTS dataset rather than just going from the LJS set to a new speaker.

As training occurred, we saved checkpoint models to be tested at a range of iterations (inference, and listen to the audio). We found that at no point after 10000 iterations on the LJS model was there any convergence and the outputs ranged from nonsensical to nonexistant. This may have been a configuration error. However, using an equivalent configuration, fine-tuning on LibriTTS speakers was successful in converging and producing a model that would generate adequate speech.

Though simply listening to the audio outputs is the most obvious way to test whether a model is converging, inspecting the mel-spectrograms and attention plots (see Fig. 5) during inference can be beneficial. It may indicate whether a model at say 200 iterations that performs well is beginning to overfit at 300 iterations, by showing weakening attention.

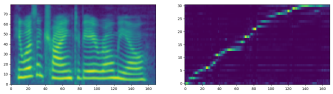

Fig. 5. Inference-time mel and attention plots

C. Tacotron2 with Waveglow training

Tacotron 2 model is trained by using a dataset that contains only the voice of one woman, which come from M-AILABS Speech Dataset, and the total number of audio files those used in this paper is 102 files. By changing the sample rate of each audio file to 22050 Hz and cleaning the raw data files, The initial code for Tacotron 2 model come from Github repository[8], the model is be able to be trained smoothly through 50 epochs using Dropout to regularize the LSTM layers, with decoder dropout rate and attention dropout rate equal to 0.1. This model is trained on uniformly sampled normalized text, with 0.005 learning rate. There is no data augmentation. With the end result of Tacotron 2 model as mel-spectogram, this result is passed to another neural model, called WaveGlow. WaveGlow is a flow-based neural network that synthesis an audio file based on Gaussian distribution by using mel-spectrogram conditioning [2].

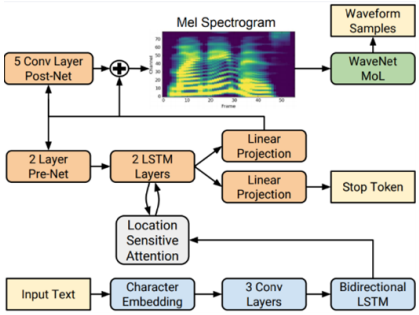

Fig. 6. Architecture of the Tacotron 2 model with WaveGlow[9]

After training Tacotron 2 model phrase with provided dataset, The audio files are generated with 10 randomly pick targets from M-AILABS Speech Dataset combining with those sentences that appear in the initial paper[1]. These are the list of sentences, which are synthesized into audio files .

The results show that these all of the sentences are synthesized into understandable audio files, with natural speech of a woman. However, at the end of these seven among all audio files contain distortions sound. This might be caused by not sufficient training time, and the limitation of the computation power.

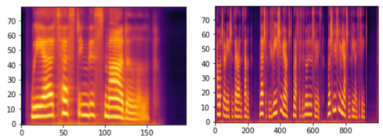

Fig. 7. Tacotron2 Mel-spectrograms during synthesis phrase

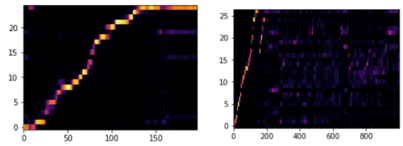

Fig. 8. Validation loss during synthesis phrase

IV. Evaluation

To evaluate the MOS of our Flowtron and Tacotron2 models we generated 10 audio samples generated from different texts for both models, along with 10 samples of real human speech, then we created a questionnaire using Google Forms and asked listeners to rate the quality of each audio snippet from 1 to 5.

Note that the test utterances are unseen by both models, as should be the case for testing.

The listeners were mainly ECS students. The MOS results are shown in Table I.

We compare these results to those from [1] in Table II, where we reference the equivalent MOS from [1] as FMOS, Human's MOS as H, and Humans's MOS from [1] as FH.

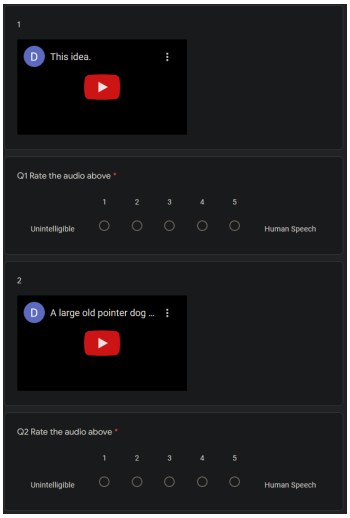

Fig. 9. Questionnaire, with a video (sound clip) to play followed by a rating 1-5

Speech Method	MOS
Human	4.970
Tacotron	3.150
Flowtron	**4.380**

TABLE I
MOS GENERATED FROM THE RESULTS OF THE QUESTIONNAIRE

Speech Method	MOS - FMOS	MOS - H	FMOS - FH
Human	+0.696	-	-
Tacotron	-0.371	-1.820	-0.753
Flowtron	**+0.715**	**-0.590**	**-0.609**

TABLE II
MOS COMPARISONS WITH FLOWTRON PAPER

From this we observe

1) Human MOS >> Human FMOS
2) Flowtron MOS >> Flowtron FMOS
3) Tacotron MOS < Tacotron FMOS
4) Flowtron MOS − Human MOS is very similar to [1]

1 and 2 and 3 indicate that our sample were generally rating much higher than the sample from [1], but that our results from Tacotron were worse than in [1]. 4 indicates the real performance of our trained Flowtron model, as it is a measure of how close our model is to human speech MOSs. We see it is quite close to that of [1] — this is important as it is an indicator of reproducibility — if not slightly overperformed.

Obviously, there are many grains of salt to be taken when comparing these results to those from [1].

- We have many less participants (not wanting to use Mechanical Turk)
  - higher variance in results
  - our sample size's opinion may not be representative of a larger population
- We trained using a smaller dataset due to issues with convergence
  - less confident audio predictions, and as such more odd-sounding parts of speech
- We did no post-processing to remove background noise
  - people may hear static and interference and rate lower because of it
- Volunteers were sent the questionnaire link without much explanation as to how to rate speech accurately
  - they often mistake 3/5 for being average sounding speech, whereas if an utterance sounds normal it should be rated 5/5 for sounding like human speech

V. CONCLUSION

Our results show that based on [1] and the code provided in the accompanying repository, it is certainly possible to reproduce equivalent results to those proposed. The instructions and documentation was lacking and as such the challenge proved difficult, we would like to note, however. A point of note is that comparatively (ironically) we found the Tacotron model's results harder to reproduce. Despite this, we regard the experiment a success.

We should also note that, as discussed, our results may need further scrutiny given more time for training models, and more resources available in order to gain a wider sample size for MOS testing.

Tests yet to be conducted include a full 'from-scratch' training of a model using the same resources as were available in [1] and the same LJS dataset to assess the claims about training and convergence time, as well as an investigation into a wider range of target voices (or languages) like gendered, heavily accented, etc. Finally, further tests into the style transfer elements of the model should be considered.

REFERENCES

[1] Valle, R., Shih, K., Prenger, R. & Catanzaro, B. Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis. (openreview.net,2020,9), https://openreview.net/forum?id=Ig53hpHxS4
[2] Prenger, R., Valle, R. & Catanzaro, B. WaveGlow: A Flow-based Generative Network for Speech Synthesis. *ArXiv:1811.00002 [cs, Eess, Stat]*. (2018,10), https://arxiv.org/abs/1811.00002
[3] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R. & Saurous, R. Tacotron: Towards End-to-End Speech Synthesis. (arXiv.org,2017), https://arxiv.org/abs/1703.10135
[4] Streijl, R., Winkler, S. & Hands, D. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*. **22** pp. 213-227 (2014,12)
[5] Kingma, D. & Dhariwal, P. Glow: Generative Flow with Invertible 1x1 Convolutions. *ArXiv:1807.03039 [cs, Stat]*. (2018,7), https://arxiv.org/abs/1807.03039
[6] Huang, C., Krueger, D., Lacoste, A. & Courville, A. Neural Autoregressive Flows. *ArXiv:1804.00779 [cs, Stat]*. (2018,4), https://arxiv.org/abs/1804.00779
[7] Imdat Solak, The M-AILABS Speech Dataset (2019), https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/
[8] Ito, K. & Seetharaman , P. NVIDIA/tacotron2. (GitHub,2020,11), https://github.com/NVIDIA/tacotron2
[9] Shen, J., Pang, R., Weiss, R., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R., Agiomyrgiannakis, Y. & Wu, Y. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. (arXiv.org,2017), https://arxiv.org/abs/1712.05884