A Lightweight Neural TTS System for High-quality German Speech Synthesis

Prachi Govalkar, Ahmed Mustafa, Nicola Pia, Judith Bauer, Metehan Yurt, †Yiğitcan Özer, Christian Dittmar

Fraunhofer IIS, Erlangen, Germany
†International Audio Laboratories Erlangen, Germany
Email: prachi.govalkar@iis.fraunhofer.de

Abstract

This paper describes a lightweight neural text-to-speech system for the German language. The system is composed of a non-autoregressive spectrogram predictor, followed by a recently proposed neural vocoder called StyleMelGAN. Our complete system has a very tiny footprint of 61 MB and is able to synthesize high-quality speech output faster than real-time both on CPU (2.55x) and GPU (50.29x). We additionally propose a modified version of the vocoder called Multi-band StyleMelGAN, which offers a significant improvement in inference speed with a small trade-off in speech quality. In a perceptual listening test with the complete TTS pipeline, the best configuration achieves a mean opinion score of 3.84 using StyleMelGAN, compared to 4.23 for professional speech recordings.

1 Introduction

The recent success of neural text-to-speech (TTS) systems combining sequence-to-sequence (S2S) spectrogram predictors with neural vocoders has led to remarkable improvements in the attainable quality of synthesized speech. In this work, we introduce a lightweight neural TTS system optimized for synthesizing natural speech output in German. It achieves a competitive Mean Opinion Score (MOS) with a tiny footprint of 61 MB and faster than real-time synthesis speed. There are three main aspects to our approach: First, our acoustic model is based on ForwardTacotron (FT) for mel-spectrogram prediction in a non-autoregressive S2S fashion. Second, we employ StyleMelGAN (SMG) [1], a novel and extremely efficient neural vocoder based on Generative Adversarial Networks (GANs). In Fig. 1, we provide a general overview of the two neural networks, which we explain in more detail in Sec. 2.1 and Sec. 3.3. Our main contribution is to show that the combination of these models yields an efficient, yet powerful neural TTS system for German language. Inspired by [2], we propose a modified Multi-band version of SMG as an additional contribution (see Sec. 3.4). The third ingredient of our system is the proprietary training dataset. We train both our acoustic model and our neural vocoder with a meticulously annotated German speech corpus comprising more than 20 hours of professional recordings from both female and male voice talents. For evaluation, we conduct a P.808 Absolute Category Rating (ACR) subjective listening test to assess the perceived speech quality. The experimental settings are detailed in Sec. 4. Our results presented in Sec. 5 corroborate the high-quality speech synthesis capabilities of our proposed system. In addition, we compare the computational requirements of the neural vocoders under test.

2 Acoustic Models

Acoustic models convert textual input (phoneme or grapheme tokens) to acoustic feature sequences, such as mel-spectrograms. Since this mapping is a S2S problem, most of the existing approaches (e.g., Tacotron [3, 4]) rely on autoregressive encoder-decoder frameworks. Commonly, there is also an attention mechanism between the encoder and decoder to estimate a temporal alignment of the tokens to the acoustic features. However, there are two known problems when using the encoder-attention-decoder mechanism for TTS: First, the generation speed is slow due to the autoregressive generation. Second, the generated speech may exhibit skipped or repeated tokens and is difficult to control on a finer level (e.g., speech rate and prosody). To alleviate these problems, non-autoregressive, duration-based acoustic models [5–9] have been proposed. As those approaches do not use attention mechanisms, the speech feature generation is both robust and controllable. Further, they are fast in inference, as the acoustic features are generated in parallel. In a return to classic TTS training principles, the authors of FastSpeech [5] introduced the combination of a duration predictor and length regulator to solve the problem of length mismatch between phoneme and mel-spectrogram sequences. The duration predictor is trained to estimate the duration of each phoneme in the input sequence. This information is then used by the length regulator to upsample (i.e., replicate) phoneme embedding vectors to match the sequence length to the desired mel-spectrogram. As a nice side-effect, the length regulator can be parameterized to control the speech rate and prosody. The complete FastSpeech model is trained jointly with the duration predictor module. The ground truth phoneme durations necessary for training are extracted from the attention matrix of a pre-trained autoregressive acoustic model. In contrast, AlignTTS [6], JDI-T [7], and EfficientTTS [8] train alignment networks jointly with their S2S models and extract durations from their alignment networks to train their duration predictors. For phoneme sequence upsampling, AlignTTS and JDI-T use the length regulator [5], whereas EfficientTTS uses a Gaussian kernel approach [8]. A completely different approach can be found in [9], where EATS combines an aligner network with the GAN-TTS [10] vocoder and trains them adversarially with a discriminator. The aligner of GAN-TTS estimates phoneme durations in time domain and calculates alignment vectors. To help the adversarial training, Soft-DTW [11] is used as the reconstruction loss between random fixed length blocks of generated and ground truth spectrograms.

2.1 ForwardTacotron

In this paper, we use FT [12] and extend it to predict multi-speaker mel-spectrograms. The top part of Fig. 1 gives more insight into the architecture, which was devised by the original author as a realization of the FastSpeech prin-

ISBN 978-3-8007-5627-8

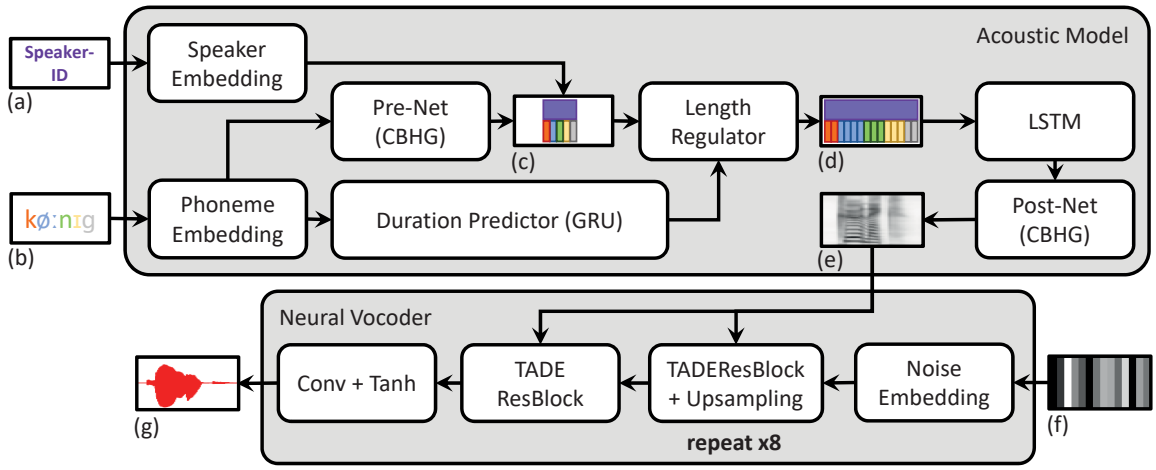

Figure 1: Simplified overview of the proposed TTS system with the ForwardTacotron acoustic model and StyleMelGAN vocoder shown as gray boxes. Neural building blocks are displayed in rounded boxes (stacking of layers shown by **repeat**), the data flowing between them are represented as smaller rectangular boxes. In particular, there are **(a)** Input speaker identity **(b)** Color-coded input phoneme sequence **(c)** Corresponding phoneme embedding sequence (color-coded), concatenated with replicated speaker embedding (purple) **(d)** Length regulated phoneme and speaker embedding sequence **(e)** Predicted mel-spectrogram **(f)** Low-dimensional noise prior **(g)** Output speech signal.

ciple with neural building blocks inspired by Tacotron [3, 4]. At the input, we provide unique speaker identifiers and phoneme sequences (including white spaces and other sentence marks). Independently trainable embedding layers convert those to hidden representations, which are concatenated after CBHG processing. In Fig. 1, the resulting speaker embeddings are visualized as purple column vectors, while the different phoneme embedding vectors are color-coded to show the correspondence to the input phoneme sequence. Prior to the pre-net, the phoneme embeddings directly serve as input to the duration predictor. While it seems counter-intuitive to make the phoneme duration prediction independent of any speaker embeddings, we still have the possibility to realize speaker-specific speech rates later in the length regulator. During FT training, the alignment matrices of a pre-trained Tacotron 2 [4] model are used to extract phoneme durations (measured in mel-spectrogram frames). This is an inexpensive alternative to having expert phoneticians provide ground truth phoneme segmentations.

3 Vocoder Models

As shown in several studies [13, 14], state-of-the-art neural vocoders outperform classical signal-processing methods [15–17] using compact speech representations, such as mel-spectrograms. So far, computationally-heavy models like WaveNet [18] and WaveGlow [19] achieved best results, while light-weight GAN models, e.g., MelGAN [20], Parallel WaveGAN [21] and Multi-band MelGAN [2] remain inferior in terms of perceptual quality. In the following sections, we describe the vocoders taken into consideration for subjective quality assessment of our TTS system (see Section 4). We give a more detailed account of SMG as this approach is a recent publication [1].

3.1 Phase Gradient Heap Integration

Phase Gradient Heap Integration (PGHI) was proposed in [16] as an efficient means for deriving phase spectrograms from Short-Time Fourier Transform (STFT) magnitude spectrograms. It uses estimates of the instantaneous frequency as well as instantaneous time (i.e., the full phase gradient) that can be approximated by the log-magnitude gradient under certain conditions. In this paper, we first convert predicted mel-spectrograms (see Sec. 2.1) to the magnitude STFT domain by frame-wise pseudo-inverse mapping from the mel-frequency scale. To remedy strong oversmoothing in the high frequency range, we multiply a noise term above 2.2 kHz. We then estimate corresponding phase spectrograms by PGHI before signal reconstruction [15].

3.2 WaveGlow

WaveGlow (WGLO) [19] uses normalizing flows to gradually transform a noise sequence into a speech waveform. It combines the flow-based approach of Glow [22] with the autoregressive WaveNet [18] architecture. The flow-based generative model provides tractability of exact log-likelihood and efficiently parallelizes both training and inference. In our earlier work [14], we identified WGLO as suitable for TTS as it yields perceptual quality scores close to WaveNet at faster than real-time inference speed on GPUs.

3.3 StyleMelGAN

The recently proposed SMG [1] is a lightweight neural vocoder for reconstructing speech from mel-spectrograms by styling a low-dimensional noise prior with the acoustic features of the target speech. It is characterized by its low complexity while still generating high-quality speech. Its main structure is a GAN comprising a generator and a discriminator network. The generator gradually transforms a noise vector into a speech signal using Temporal Adaptive DEnormalization (TADE), a technique first used in image processing [23]. This is done by upsampling the noise vector through a series of building blocks called TADE residual blocks (TADEResBlocks) as shown in the lower part of Fig. 1. These TADEResBlocks mainly consist of two TADE layers which are conditioned on the mel-spectrogram. As detailed in Fig. 2, each TADE layer ap-

ISBN 978-3-8007-5627-8

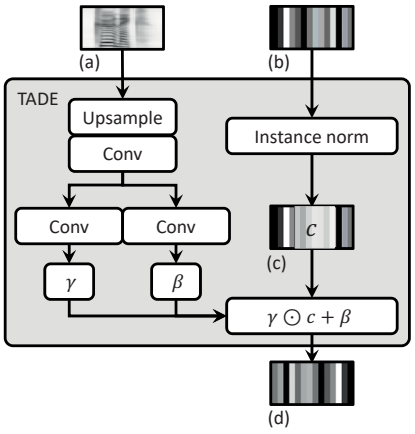

Figure 2: TADE layer architecture of StyleMelGAN. The learned modulation parameters γ and β have the same dimensionality as the normalized activation c. The symbol $\odot$ indicates the pointwise multiplication. **(a)** Conditioning mel-spectrogram **(b)** Input activation **(c)** Normalized input activation **(d)** Output activation.

plies a linear modulation of the upsampled mel-spectrogram to the input vector c through the trainable modulation parameters γ and β. In each block, c is normalized using instance norm [24] before being subjected to point-wise multiplication with γ and translation by β. The output of the TADEResBlock is computed from the sequence of TADE blocks and a skip connection from the input vector. The generator network is constructed by stacking a combination of TADEResBlocks and upsampling layers (upsample by factor two) eight times, a final TADEResBlock and a convolutional layer with tanh non-linearity. The training of the generator is guided by an adversarial loss which is computed by the discriminator network. For the discriminator, an ensemble of four discriminator networks (each based on [2]) is used. The input of the discriminators are random windows sliced from the target speech signal, similar to [10]. These random segments are then decomposed by Pseudo Quadrature Mirror Filter-bank (PQMF) [25], in order to analyze the frequency bands of the signal. Both the length of the speech segments and the number of subbands for the PQMF differ between the discriminators in the ensemble (512/1024/2048/4096 samples per segment, 1/2/4/8 subbands for PQMF). Following the PQMF, there is a convolutional layer and a sequence of three blocks consisting of a 1D convolution, a LeakyReLU and a downsampling operation. Finally, there is a 1D convolution, a LeakyReLU and a convolutional layer. For training the GAN, first only the generator is pretrained using the spectral reconstruction loss. Then, the generator and the discriminator are trained together which helps to obtain more natural speech signals. For the loss of the generator, the sum of the adversarial loss from the discriminator and the spectral reconstruction loss is used in order to prevent adversarial artifacts. Due to its low complexity, SMG is capable of synthesizing speech signals with 22.05 kHz more than 50 times faster than real-time on GPU. In our earlier paper [1], the quality of the synthesized speech was evaluated using Fréchet scores and listening tests. By computing the conditional Fréchet Deep Speech Distance [10, 26], we showed that SMG outperforms other vocoders. A MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) listening test for copy-synthesis showed that SMG outperforms other vocoders by about 15 MUSHRA points. To also evaluate SMG in a TTS scenario, a P.800 [27] listening test was performed where it achieved a MOS of 4.00 ± 0.06, outperforming other models.

Condition	Description
FT + PGHI	FT + Phase Gradient Heap Integration
FT + WGLO	FT + WaveGlow
FT + MBSMG	FT + Multi-band StyleMelGAN
FT + SMG	FT + StyleMelGAN
REF	Reference speech recordings

Table 1: Synthesis conditions under test. Here, FT stands for ForwardTacotron.

3.4 Multi-band StyleMelGAN

Based on SMG, we developed Multi-band StyleMelGAN (MBSMG). As shown in Fig. 3, the generator of MBSMG is similar to the generator of SMG. In contrast to Fig. 1, we depict the signal flow going from left to right for the sake of clarity. As a main difference, the last two TADEResBlocks are not followed by an upsampling layer resulting in a tensor with one quarter the size compared to the SMG case. Thus, the final convolutional layer reduces the number of channels to 4. These 4 subbands are then combined to the final speech signal by a PQMF synthesis layer. So instead of directly computing the complete speech signal, 4 subbands are computed and combined by a PQMF. This modified structure of the generator leads to a higher synthesis speed compared to the generator structure of SMG. The reason for this speed-up is that especially the last two TADEResBlocks of SMG are computationally expensive which is improved in MBSMG by reducing the data dimensions in these layers.

4 Experiment

For our experiments, we compared different versions of our proposed TTS system by keeping the same acoustic model (i.e., FT) while exchanging the vocoder models. The 5 different test conditions are explained in Tab. 1.

4.1 Corpus and Audio Processing

We use a proprietary dataset which comprises 20 hours of speech recordings performed by two native German speakers in studio recording conditions. 48% of the dataset is spoken by a professional female voice actor and the remaining part by a professional male voice actor. The original recordings are encoded as 32-bit mono PCM with a sampling rate of 48 kHz. For training, all speech signals were resampled to 22.05 kHz. Further preprocessing involved DC offset removal and max normalization. Mel-spectrograms with 80 bands were extracted for each speech signal (using 46.4 ms block size and 11.6 ms hop size) in the frequency range 0 kHz to 8 kHz. For WGLO, we used a pretrained model [1] for warmstarting and then continued to train with our proprietary dataset. Both SMG and MBSMG were trained from scratch using our dataset.

[1] https://ngc.nvidia.com/catalog/models/nvidia:waveglow_ljs_256channels

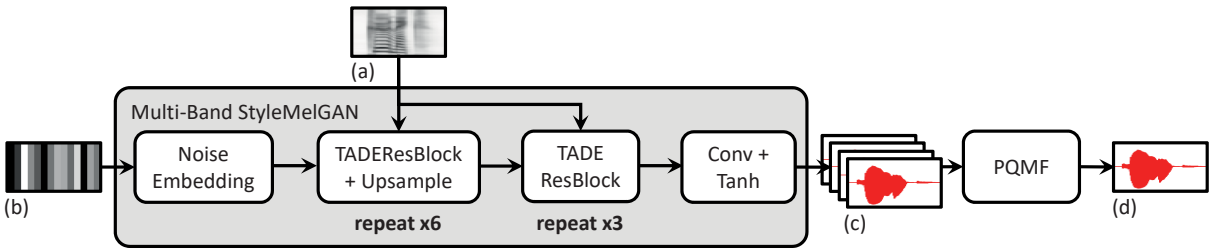

Figure 3: Generator architecture of Multi-band StyleMelGAN (MBSMG). **(a)** Conditioning mel-spectrogram **(b)** Low-dimensional noise prior **(c)** Generated speech sub-band signals **(d)** Generated speech signal.

Condition	Average	Male	Female
FT + PGHI	1.3 ± 0.04	1.17 ± 0.04	1.41 ± 0.07
FT + WGLO	2.72 ± 0.07	2.26 ± 0.09	3.17 ± 0.1
FT + MBSMG	3.38 ± 0.07	3.21 ± 0.1	3.54 ± 0.09
FT + SMG	3.84 ± 0.06	3.79 ± 0.09	3.9 ± 0.09
Reference	4.23 ± 0.06	4.32 ± 0.08	4.13 ± 0.09

Table 2: MOS-scores with 95% confidence intervals for male and female speakers along with average scores.

Condition	Spect. type	Model Size (in MB)[3]	#Param. (in M)[4]	RTF	
				CPU	GPU
FT + PGHI	Linear	-	-	15.48	39.68
FT + WGLO	Mel	170	86.3	0.57	8.75
FT + MBSMG	Mel	15	3.85	4.35	61.27
FT + SMG	Mel	15	3.85	2.55	50.29

Table 3: Model size, parameter count and real-time factor. Here, we report a combined RTF of acoustic model and neural vocoder, the higher the better. The inference speed was measured on CPU (Intel Core i7-8700K 3.70 GHz) and a single GPU (NVIDIA GeForce GTX 1080 Ti).

4.2 Setup and Participants

The subjective evaluations for the TTS pipelines were done using a P.808 ACR listening test [28] by crowdsourcing.[2] The test was performed by 36 German native speakers with an average age of 35.75 years (±11.26). 15 participants were recruited through Amazon Mechanical Turk (MTurk) [29]. On average, the participants took 15 minutes to finish the test. WebMUSHRA [30], a popular framework for conducting web-based listening tests, was customized for a 5-point MOS scale. This framework was hosted on a private domain with the help of Amazon Web Services [31]. In accordance with the P.808 recommendations, the MTurk workers were tested for native-level fluency, subjected to multiple gold standard questions during the test and had to undergo a hearing screening task to ensure wearing headphones. The subjective evaluation was divided into two phases: Training and Testing. The training phase allowed the participants to familiarize themselves with the audio quality provided by different systems under test and the user interface of WebMUSHRA. We selected two items for this phase and ten for the testing phase. These items were synthesized both in male and female voices for all the test conditions (see Tab. 1), leading to a total of 120 items. The participants rated these items independently, based on naturalness and intelligibility, hence there was no direct comparison between the conditions while assigning scores.

5 Results and Conclusion

Tab. 2 compares the obtained MOS-scores of all systems under test. The results clearly show that the combination of FT + SMG outperforms all the other systems and can generate high-quality speech with a MOS of 3.84. In line with the expectations, neural vocoders like SMG, MBSMG and WGLO have an obvious advantage over phase reconstruction based methods. It is worth noting that both SMG and MBSMG achieved much better scores for synthesizing the male voice in comparison to WGLO. On closer inspection, we found that they yielded more clarity and coherence in the pitched parts of male speech, whereas WGLO sometimes tended to produce noisy and trembling sound. The current results are not directly comparable to our earlier findings in [1], since we used a German dataset and FT as acoustic model.

In order to compare the model complexity, the memory consumption, the number of trainable parameters, and the real-time factors are summarized in Tab. 3. Although the model size remains the same for FT + SMG and FT + MBSMG, the latter has a noticeable increase in real-time factor at the cost of a dip in speech quality.

In summary, our proposed TTS system yields competitive results in comparison to other state-of-the-art TTS methods. Future work will be directed towards further optimization of the computational requirements.

6 Acknowledgements

Parts of this work have been supported by the SPEAKER project (FKZ 01MK20011A), funded by the German Federal Ministry for Economic Affairs and Energy. In addition, this work was supported by the Free State of Bavaria in the DSAI project. The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

[2] Speech items used in the listening test are available at: https://www.audiolabs-erlangen.de/resources/NLUI/2021-FT-SMG-TTS

[3] Only vocoder model sizes are displayed for comparison purposes. Since the same FT model is used for all TTS systems, its model size remains the same (46 MB).

[4] Similarly, the number of parameters are only mentioned for the vocoder models. The FT model has 23.94 M parameters.

References

[1] A. Mustafa, N. Pia, and G. Fuchs, "StyleMelGAN: An Efficient High-Fidelity Adversarial Vocoder with Temporal Adaptive Normalization," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Toronto, ON, Canada), pp. 6034–6038, June 2021.

[2] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band MelGAN: Faster Waveform Generation for High-Quality Text-to-Speech," in *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*, (Virtual), pp. 492–498, January 2021.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," in *Proc. of Interspeech*, (Stockholm, Sweden), pp. 4006–4010, August 2017.

[4] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on Mel-Spectrogram Predictions," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Calgary, AB, Canada), pp. 4779–4783, April 2018.

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

[6] Z. Zeng, J. Wang, N. Cheng, T. Xia, and J. Xiao, "AlignTTS: Efficient Feed-Forward Text-to-Speech System without Explicit Alignment," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Barcelona, Spain), pp. 6714–6718, May 2020.

[7] D. Lim, W. Jang, G. O, H. Park, B. Kum, and J. Yoon, "JDI-T: Jointly trained Duration Informed Transformer for Text-To-Speech without Explicit Alignment," *CoRR*, vol. 2005.07799, 2020.

[8] C. Miao, S. Liang, Z. Liu, M. Chen, J. Ma, S. Wang, and J. Xiao, "EfficientTTS: An Efficient and High-Quality Text-to-Speech Architecture," *CoRR*, vol. 2012.03500, 2020.

[9] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, "End-to-End Adversarial Text-to-Speech," *CoRR*, vol. abs/2006.03575, 2020.

[10] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High Fidelity Speech Synthesis with Adversarial Networks," *CoRR*, vol. abs/1909.11646, 2019.

[11] M. Cuturi and M. Blondel, "Soft-DTW: a Differentiable Loss Function for Time-Series," in *Proc. of the International Conference on Machine Learning (ICML)*, (Sydney, NSW, Australia), pp. 894–903, August 2017.

[12] C. Schäfer, "ForwardTacotron." `https://github.com/as-ideas/ForwardTacotron`, 2020. Accessed: 2021-05.

[13] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, "A Comparison of Recent Waveform Generation and Acoustic Modeling Methods for Neural-Network-Based Speech Synthesis," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Calgary, AB, Canada), pp. 4804–4808, April 2018.

[14] P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, "A Comparison of Recent Neural Vocoders for Speech Signal Reconstruction," in *Proc. of the ISCA Speech Synthesis Workshop*, (Vienna, Austria), pp. 7–12, September 2019.

[15] D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[16] Z. Pruša, P. Balázs, and P. L. Søndergaard, "A Noniterative Method for Reconstruction of Phase from STFT Magnitude," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1154–1164, 2017.

[17] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "A Comparison Between STRAIGHT, Glottal, and Sinusoidal Vocoding in Statistical Parametric Speech Synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1658–1670, 2018.

[18] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *Proc. of the ISCA Speech Synthesis Workshop*, (Sunnyvale, CA, USA), pp. 125–125, September 2016.

[19] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Brighton, UK), pp. 3617–3621, May 2019.

[20] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, (Vancouver, BC, Canada), pp. 14881–14892, December 2019.

[21] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Barcelona, Spain), pp. 6199–6203, May 2020.

[22] D. P. Kingma and P. Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions," in *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, (Montréal, QC, Canada), pp. 10215–10224, December 2018.

[23] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic Image Synthesis with Spatially-Adaptive Normalization," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Long Beach, CA, USA), pp. 2337–2346, June 2019.

[24] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," *CoRR*, vol. abs/1607.08022, 2016.

[25] T. Q. Nguyen, "Near-perfect-reconstruction pseudo-QMF banks," *IEEE Transactions on Signal Processing*, vol. 42, no. 1, pp. 65–76, 1994.

[26] A. A. Gritsenko, T. Salimans, R. v. d. Berg, J. Snoek, and N. Kalchbrenner, "A Spectral Energy Distance for Parallel Speech Synthesis," *CoRR*, vol. abs/2008.01160, 2020.

[27] I. Rec, "P. 800: Methods for subjective determination of transmission quality," *International Telecommunication Union, Geneva*, vol. 22, 1996.

[28] I. Rec, "P. 808: Subjective evaluation of speech quality with a crowdsourcing approach," *International Telecommunication Union, Geneva*, 2018.

[29] "2005-2018, Amazon Mechanical Turk, Inc.." `https://www.mturk.com/`. Accessed: 2021-05-25.

[30] M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre, "Towards the Next Generation of Web-based Experiments: A Case Study Assessing Basic Audio Quality Following the ITU-R Recommendation BS.1534 (MUSHRA)," in *Proc. of the Web Audio Conference (WAC)*, (Paris, France), 2015.

[31] "2020, Amazon Web Services, Inc.." `https://aws.amazon.com/`. Accessed: 2021-05-25.

ISBN 978-3-8007-5627-8