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

Introduction

Our lightweight neural TTS system is optimized for synthesizing natural speech output in German and has 3 main aspects:

- **Acoustic model**
  - Textual input (phoneme/grapheme tokens) → acoustic feature sequences (mel-spectrograms)
  - Our implementation based on ForwardTacotron[1] (FT), predicts mel-spectrograms in a non-autoregressive seq-2-seq fashion
- **Vocoder model**
  - Acoustic feature sequences → time-domain audio signals
  - We use StyleMelGAN [1] (SMG), a novel and extremely efficient neural vocoder based on Generative Adversarial Networks
- **Proprietary speech corpus for training both models**
  - 20 hrs of professional speech recordings by 2 native German speakers

Inspired by [2], we propose a modified Multi-band version of SMG (MBSMG) as an additional contribution.

Experiment

We compare different versions of TTS systems by keeping the same acoustic model (i.e., ForwardTacotron) while exchanging the vocoder models. The vocoder models are as follows:

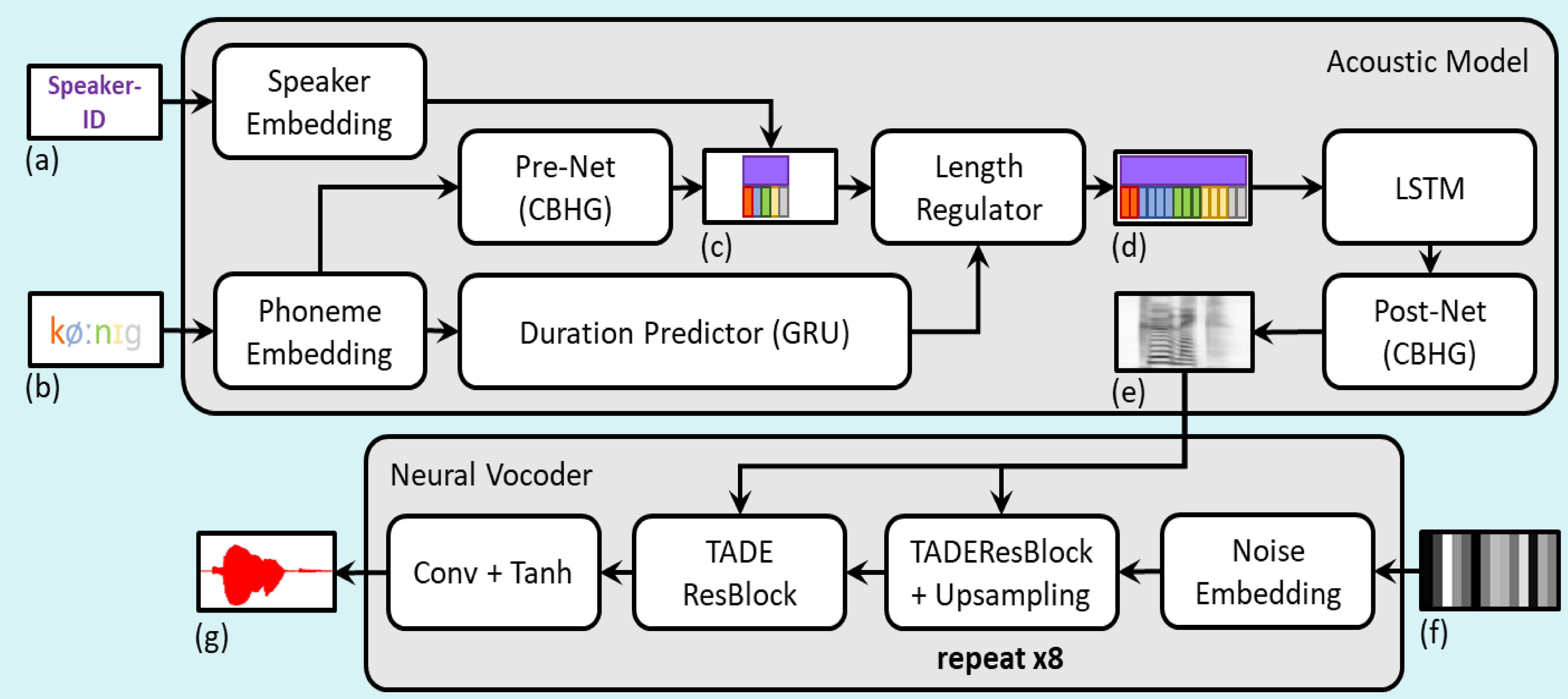
- Phase Gradient Heap Integration (PGHI) [3]
- WaveGlow (WGLO) [4]
- StyleMelGAN (SMG)
- Multi-band StyleMelGAN (MBSMG)

MBSMG synthesizes speech subbands, combined by a Pseudo Quadrature Mirror Filter-bank [5], leads to higher synthesis speed (see RTF in Table 2).

Setup

- Synthesized speech signals preprocessed by applying DC offset removal and max normalization, 80-band mel-spectrograms (freq. range 0-8 kHz)
- SMG and MBSMG trained from scratch using our dataset
- WGLO finetuned, warmstarted using pretrained model[2]
- P.808 [6] ACR listening test
- 36 German native speakers,15 fromAmazon Mechanical Turk[3]
- Web-based listening tests using WebMUSHRA [7]

[1] C. Schäfer, "ForwardTacotron." https://github.com/as-ideas/ForwardTacotron, 2020.
[2] https://ngc.nvidia.com/catalog/models/nvidia:waveglow_ljs_256channels
[3] https://www.mturk.com/
[4] Model sizes and number of parameters are displayed only for vocoder models since acoustic model remains same
[5] Inference speed on CPU (Intel Core i7-8700K 3:70 GHz) and a single GPU (NVIDIA GeForce GTX 1080 Ti)



a) Input speaker identity
b) Input phoneme sequence
c) Corresponding phoneme embedding sequence, concatenated with replicated speaker embedding
d) Length regulated phoneme and speaker embedding sequence (here, phoneme durations obtained using alignment matrices of pretrained Tacotron 2 [8] model)
e) Predicted mel-spectrogram
f) Low-dimensional noise prior
g) Output speech signal

Figure 1: Simplified overview of the proposed TTS system with the ForwardTacotron acoustic model and StyleMelGAN vocoder

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

Condition	Spect. Type	Model Size (in MB)[4]	#Para. (in M)[4]	RTF[5]	
				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 2: Model size, parameter count and real-time factor (combined from both acoustic model and neural vocoder).

Results

- FT + SMG outperforms all the other systems, generates high-quality speech (MOS -3.84) (see Table 1)
- SMG and MBSMG achieve higher scores for synthesizing male voice in comparison to WGLO, improved clarity and coherence in the pitched parts
- SMG and MBSMG are extremely lightweight in comparison to WGLO, achieve high inference speeds on CPU and GPU

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.

Checkout audio samples used in the listening test by scanning the QR-code or **click here**. Headphones preferred.

[1] A. Mustafa et al., "StyleMelGAN: An Efficient High-Fidelity Adversarial Vocoder with Temporal Adaptive Normalization," in Proc. ICASSP 2021.
[2] G. Yang et al., "Multi-band MelGAN: Faster Waveform Generation for High-Quality Text-to-Speech," in Proc. IEEE SLT, 2021
[3] Z. Pruša et al., "A Noniterative Method for Reconstruction of Phase from STFT Magnitude, "IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, 2017.
[4] R. Prenger et al., "WaveGlow: A Flow-based Generative Network for Speech Synthesis," in Proc. ICASSP 2019.
[5] T. Q. Nguyen, "Near-perfect-reconstruction pseudo-QMF banks," IEEE Transactions on Signal Processing, 1994.
[6] I. Rec, "P. 808: Subjective evaluation of speech quality with a crowdsourcing approach," International Telecommunication Union, Geneva, 2018.
[7] M. Schoeffler et al., "Towards the Next Generation of Web-based Experiments: A Case Study Assessing Basic Audio Quality Following the ITUR Recommendation BS.1534 (MUSHRA)," in Proc. WAC 2015.
[8] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on Mel-Spectrogram Predictions," in Proc. ICASSP 2018.