Neural Speech Synthesis in German

Based on Tacotron 2 and Multi-Band MelGAN

Johannes Wirth, Pascal Puchtler, René Peinl

Contact: johannes.wirth.3@iisys.de

Hochschule Hof
University of Applied Sciences

Hochschule Hof
University of Applied Sciences

Johannes Wirth

- Research fellow at Institute for Information Systems (iisys) at Hof University of Applied Sciences, Germany
- M. Sc. In Computer Science on “Evaluating Transformer-Based Language Models for German Speech Recognition” at Hof University of Applied Sciences
- Research Focus: NLP and NLU based on neural networks for the German language

Hochschule Hof
University of Applied Sciences

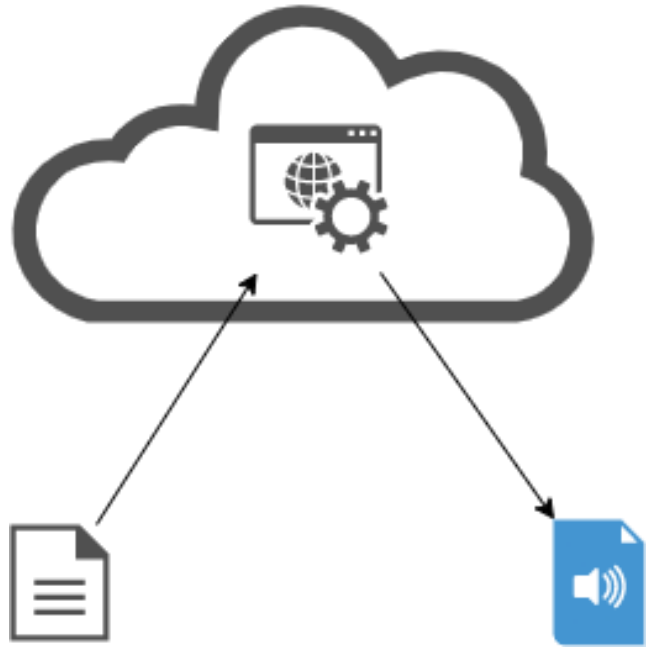
Motivation

Introduction - SOTA in Text To Speech

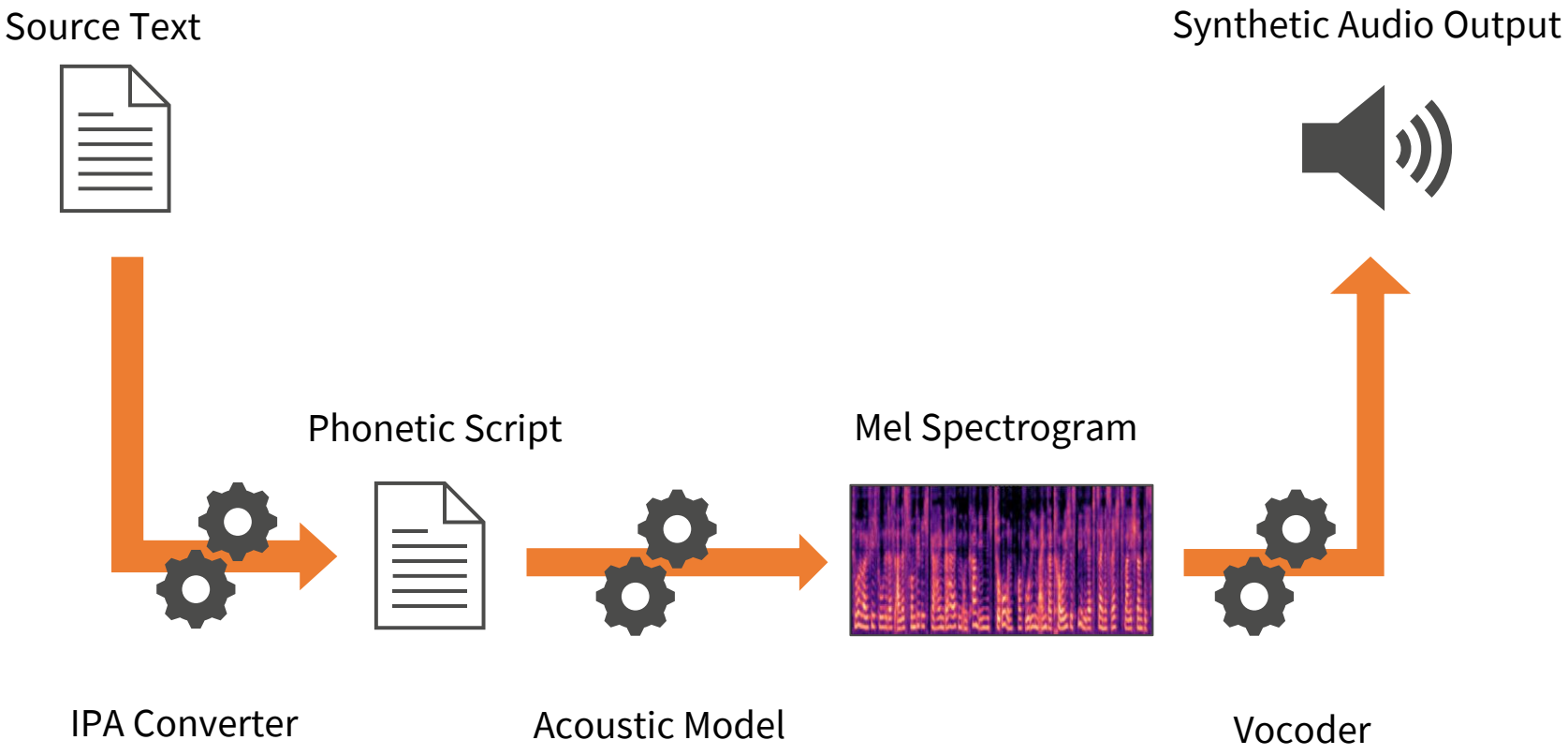
1. Dataset Selection
2. Model Selection
3. Model Training
4. Empirical Analysis
5. Conclusion
6. Future Work

1. Motivation

- Text-To-Speech (TTS) architectures are mostly only evaluated using English datasets
- Trained models are rarely published for free use
- Companies intending to use TTS rely on large-scale SAAS providers
- Challenges in training models for the German language have scarcely been investigated



Introduction - SOTA in Text To Speech



2. Dataset Selection

Requirements loosely derived from LJSpeech:

- Minimum length of aligned audio-transcript pairs of >20 hours
- Text normalization
- Preferable sampling rate of 22.05 kHz

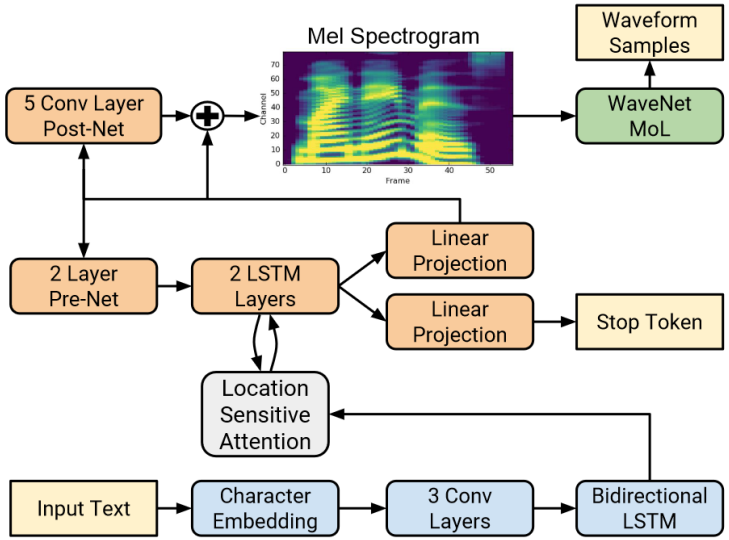
Further Dataset Processing:

- Punctuation reduction
- Phonemization

Dataset	Speaker	Sampling Rate	Length
HUI Audio Corpus	Bernd Ungerer (m)	22 kHz	97 h
	Hokuspokus clean (f)	22 kHz	27 h
	Hokuspokus full (f)	22 kHz	43 h
Thorsten neutral	Thorsten Müller (m)	22 kHz	23 h
M-AILABS	Eva K (f)	16 kHz	29 h
	Karlsson (m)	16 kHz	40 h

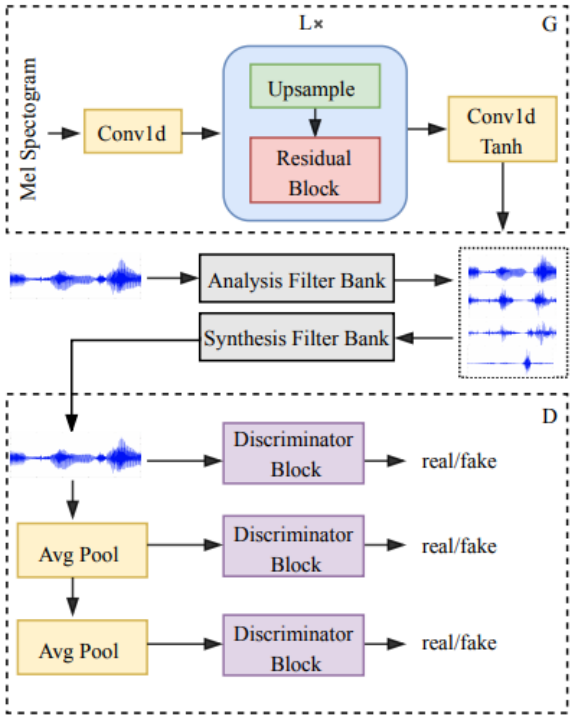
3. Model Selection

Acoustic Model (Tacotron 2)*



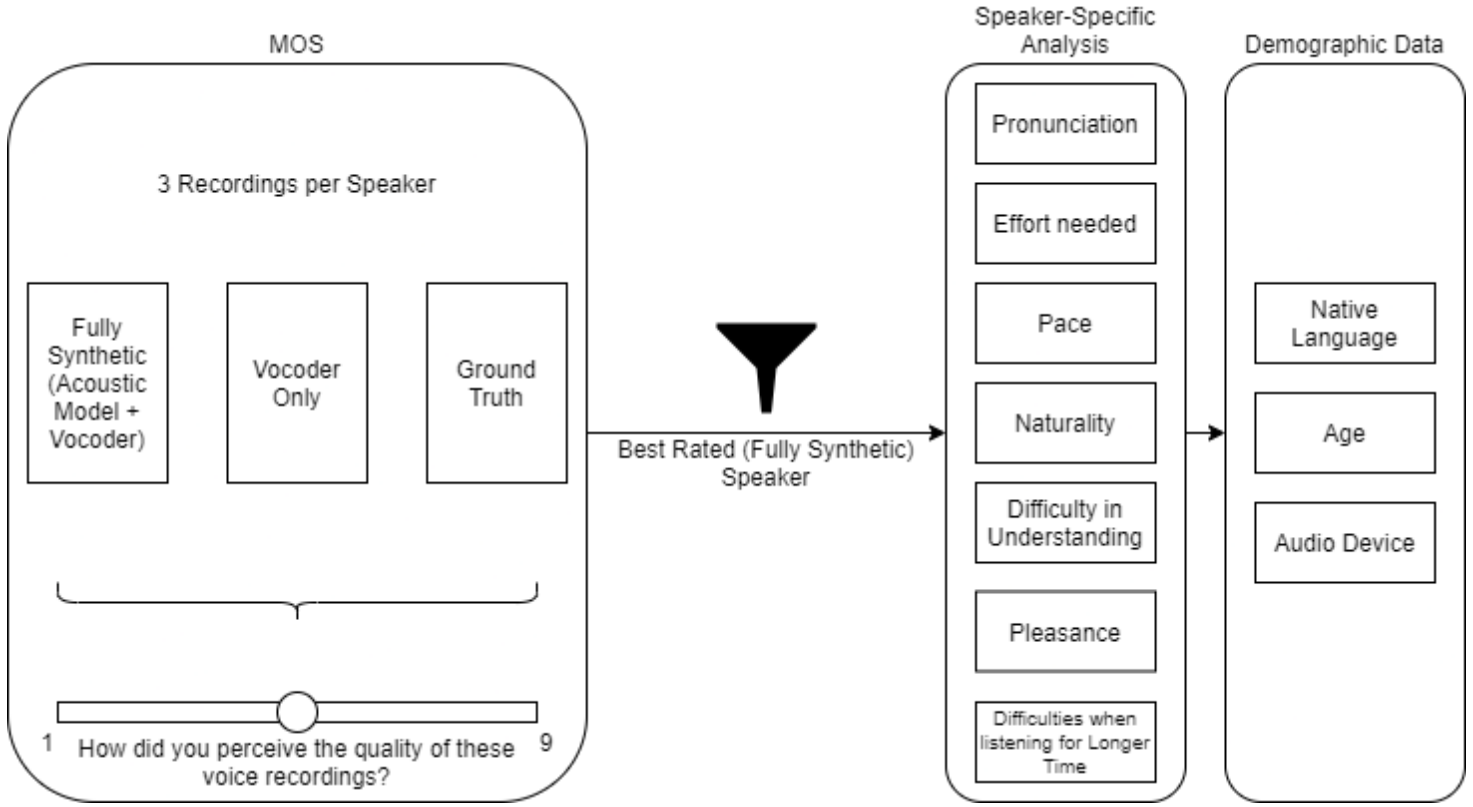
Modifications: Guided Attention Loss, AMSGrad

Vocoder (Multi-Band MelGAN)**



*J. Shen et al., "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4779–4783.
**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," arXiv preprint arXiv:2005.05106, 2020.

4. Empirical Analysis – Questionnaire Design



4. Empirical Analysis - Results

Dataset	Speaker	Synth	Δ GT	Vocoder	GT
HUI Audio Corpus	Bernd Ungerer	3.74	0.51	3.75	4.25
	Hokuspokus clean	2.98	1.29	x	4.27
	Hokuspokus full	2.88	1.39	3.60	4.27
Thorsten neutral	Thorsten Müller	3.49	0.50	3.78	3.99
M-AILABS	Eva K	2.13	1.60	3.33	3.72
	Karlsson	2.96	1.18	3.76	4.14

Hochschule Hof
University of Applied Sciences

4. Empirical Analysis – Results - Speaker-Specific Analysis

Speaker	Votes	Q1 (5.0)	Q2 (5.0)	Q3 (0.0)	Q4 (5.0)	Q5 (5.0)	Q6 (5.0)	Q7 (5.0)
Bernd Ungerer	54	3.6	4.4	-0.2	4.0	4.1	3.9	3.5
Thorsten Müller	14	3.7	4.3	-0.2	3.1	4.0	3.5	3.0
Hokuspokus Clean	3	3.2	4.2	±0	3.2	4.2	3.3	3.5
Hokuspokus Full	23	3.0	4.1	-0.3	3.3	3.6	3.6	3.0

- Pronunciation(Q1)
- Effort needed (Q2)
- Pace (Q3)
- Naturality(Q4)
- Difficulties in Understanding(Q5)
- Pleasance (Q6)
- Difficulties when listening for longer time(Q7)

5. Conclusion

- An initial benchmark was set for upcoming TTS models in German
- Trained models show comparable performance to SOTA in English and are published for free use
- Certain linguistic aspects and dataset properties influence the quality of trained models significantly
- Synthetic voices still lack certain elements of naturality
- Further improvable aspects regarding
    1. Dataset quality
    2. Model hyperparameters
    3. Data preprocessing

  were identified

6. Future Work

- Investigation of resource-efficient architectures for use on edge devices (smart speaker)
- Further adaptions of data, hyperparameters and phonemization
- Training of additional vocoders
- Identification of crucial linguistic aspects in training data

Hochschule
Hof

University of
Applied Sciences

Johannes Wirth, M. Sc.
Institute of Information Systems at Hof University
95028 Hof
Alfons-Goppel-Platz 1
Phone +49 9281 409-6201
Johannes.wirth.3@iisys.de
www.iisys.de