Neural Speech Synthesis in German

Based on Tacotron 2 and Multi-Band MelGAN

Johannes Wirth, Pascal Puchtler, René Peinl
Research Group System Integration
Hof University of Applied Sciences
Alfons-Goppel-Platz 1, 95028 Hof, Germany
e-mail: Johannes.Wirth.3@iisys.de, Pascal.Puchtler@iisys.de, Rene.Peinl@iisys.de

***Abstract*—While many speech synthesis systems based on deep neural networks are thoroughly evaluated and released for free use in English, models for languages with far less active speakers like German are scarcely trained and most often not published for common use. This work covers specific challenges in training text to speech models for the German language, including dataset selection and data preprocessing, and presents the training process for multiple models of an end-to-end text to speech system based on a combination of Tacotron 2 and Multi-Band MelGAN. All model compositions were evaluated against the mean opinion score, which revealed comparable results to models in literature that are trained and evaluated on English datasets. In addition, empirical analyses identified distinct aspects influencing the quality of such systems, based on subjective user experience. All trained models are released for public use.**

Keywords: Text-To-Speech; German; Tacotron 2; Multi-Band MelGAN.

I. Introduction

The quality of speech synthesis or Text To Speech (TTS) systems has leaped since deep neural networks are being leveraged. Whereas such systems acted as a niche technology a few years ago, today every voice assistant and a large number of car models are equipped with their own, manufacturer-specific, synthetic but increasingly natural-sounding voices. However, smaller companies interested in using TTS in their products or services mostly have to rely on large-scale software providers, or alternatively, freely available models as investments in in-house solutions would often be financially unfeasible.

Since state-of-the-art models with permissive licenses exist almost exclusively for English, several model compositions based on Tacotron 2 [1] and Multi-Band MelGAN [2] were trained for the German language and published for free use. This work describes the processes that were carried out to train these neural networks and provides a corresponding evaluation based on the Mean Opinion Score (MOS), setting an initial benchmark for future systems in German. The described models and results are part of the development of a smart speaker system.

The rest of the paper is structured as follows. In Section II, state-of-the-art of deep neural networks used for TTS are presented and available datasets for German TTS are reviewed in section 3. In section 4, key learnings from training of selected network models are described further. Section 5 describes how the evaluation of synthetic voices was implemented and presents the results, which are interpreted in a subsequent discussion in section 6 and put into perspective by limitations in section 7. Lastly, the work is concluded with a summary and an outlook in section 8.

II. BACKGROUND

Most state-of-the-art systems for speech synthesis based on neural networks consist of two components: an acoustic model and a vocoder. The acoustic model generates an intermediate representation called mel spectrogram from input characters or phonemes, while the vocoder converts this representation into a final audio signal. The following subsections describe the general principles of operation of both components in more detail and presents several architectures. An overview of model compositions already evaluated in literature is given in TABLE II.

A. *Acoustic Model*

Acoustic modelling defines the task of encoding an input sequence of characters to a hidden representation and the subsequent prediction of mel spectrogram frames per time step. The formerly common models for mel spectrogram generation based on Hidden Markov Models (HMMs) [3] have been increasingly replaced by approaches based on deep learning in recent years. In particular, Tacotron [4] and its successor Tacotron 2 [1] have led to a dramatic increase of quality in speech synthesis research. While Tacotron still uses a Griffin-Lim vocoder as a second stage, only reaching a MOS of 3.82, Tacotron 2 succeeds in achieving a MOS value of 4.53, which is very close to the value of human speakers (4.58), by using a continuous deep learning-based process. For the latter, a modified version of WaveNet [5] was used as a vocoder.

While Tacotron is based on Recurrent Neural Networks (RNNs), which are commonly used for speech synthesis, Transformer TTS [6] successfully applied the transformer architecture [7], which became well-known from the domain of natural language processing with models such as BERT [8], to speech synthesis, achieving similar or slightly better scores than Tacotron 2. Transformer TTS [9] achieves a MOS value of 4.39 compared to 4.44 of human speakers and is thus on par with Tacotron 2.

Autoregressive models such as Tacotron 2 and Transformer TTS achieve state-of-the-art quality but can hardly be parallelized, leading to longer processing times. A few minutes of audio quickly take hours to generate [10]. Therefore, most of the research in 2019 and 2020 has focused on exploring architectures that are significantly faster and provide similarly good MOS values, rather than continuing to work on even better speech quality. Both Tacotron2 and TransformerTTS also incorporate certain attention mechanisms, which can lead to word omissions or even repetitions in outputs.

Non-autoregressive models can be further categorized into those using knowledge distillation like FastSpeech [10] and others utilizing differing technologies. Flow-TTS [11] and Glow-TTS [12] are examples for the latter. Interestingly, while many of the more recent publications presenting non-autoregressive models claim to be better than Tacotron 2 in a direct comparison, none of them were able to achieve comparably good MOS values close to the ground truth. Parallel Tacotron [13], Flow-TTS [11] and Fastpitch [14] are closest with MOS values above 4.0 and less than 0.5 worse than the ground truth.

B. Vocoder

Neural vocoders receive a mel spectrogram and predict audio signal frames for each spectrogram frame. A mel spectrogram can be generated directly from an audio file, as opposed to acoustic models, requiring audio-transcript-pairs. Therefore, it is comparably easy to generate training data, which results in a broad selection of well performing vocoders that can produce high quality audio hardly distinguishable from real human voices. The main reference is WaveNet [5], which achieved 4.21 on the MOS scale from 1 to 5 in the original publication [5] and 4.53 MOS in a later publication [1]. This is very close to the ground truth of 4.58 and still the state-of-the-art reference value up until now. Since WaveNet is autoregressive, it is both comparably slow and requires significant resources. To compensate these weak points, several alternatives have been suggested.

Parallel WaveNet [15] uses knowledge distillation to derive a much faster network from WaveNet in a student-teacher manner. It can generate 20s of audio in 1s (real-time factor RTF 0.05), whereas WaveNet requires 1,000s to generate 20s of audio (RTF 50).

WaveGlow [16] is a representative of flow-based networks, which can be parallelized well in contrast to auto-regressive networks like WaveNet. It achieves RTF 0.04 on an Nvidia Tesla V100 GPU. It is also commonly implemented as acoustic model, i.e., in [12], [17], [18].

Multi-Band MelGAN [2] is also worth mentioning, being based on a different approach. Its architecture utilizes a Generative Adversarial Network (GAN) and achieved a MOS of 4.34 in empirical analysis. However, this was achieved for the Chinese language instead of English and is therefore not directly comparable.

Best results based on the popular LJspeech dataset [19] are reported by Hifi-GAN [20] and WaveGrad [21] with 4.36 and 4.55 respectively. The latter is identical to the ground truth MOS value.

Finally, WaveRNN [22] achieves MOS 4.46 and is therefore the closest competitor to WaveNet and WaveGrad.

III. DATASETS

The selection of suitable datasets was based on metadata from LJSpeech. Strict criteria for the minimum length of audio-transcript pairs (>20 hours) and text normalization (no leftover digits or symbols) were set. The sampling rate of 22.05kHz was not considered to be a hard criterion, merely regarded preferable, so not to further reduce the scope of the already limited number of existing datasets.

Selected datasets were further processed in preparation of the subsequent training processes.

A. Selection

Besides the acoustic model and vocoder, the quality and quantity of the dataset used for training are the main factors influencing the quality of the resulting synthetic voice. The following datasets were evaluated regarding their suitability and partially selected for subsequent model training. The final selection of datasets is presented in TABLE **I**.

1) M-AILABS

The M-AILABS speech dataset is based on data from LibriVox [23], a platform providing free audio books by voluntary, mostly amateur speakers, and consists of five single speaker datasets. Their durations range from 19h to 68h of speech and respective texts. Despite a comparatively low sampling rate of 16kHz for each recording, two speakers, Karlsson (male, 40h) and Eva K (female, 29h) were chosen for model training. Ramona (female, 68h) was discarded due to her subjectively unpleasant voice.

2) Thorsten Voice

Specifically created for the creation of TTS applications, the Thorsten neutral dataset consists of more than 23 hours of audio-transcript pairs from a single male voice, recorded with a sampling rate of 22.05kHz [24]. It was first released in March 2021 and, to the authors' knowledge, has not been evaluated in any scientific publication yet.

3) HUI Audio Corpus

Similar to M-AILABS, the recently released HUI audio corpus [25] also consists of freely available audio data from LibriVox and transcripts from gutenberg.org [26], but provides a much larger quantity of audio-transcript pairs per speaker and a higher sampling rate of 22.05kHz. The speakers Bernd Ungerer (male, 97h) as well as Hokuspokus full (female, 43h) and Hokuspokus clean (female, 27h; subset of Hokuspokus full, containing less noise) were chosen for model training.

TABLE I. DATASETS USED FOR FURTHER PROCESSING.

Dataset	Speaker	Sampling Rate	Hours
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

B. Further Processing

To reduce the range of phrases and punctuation marks acoustic models receive as input, transcript sentences of all datasets were filtered and adjusted using several mechanisms. Also, since phoneme-based models generally perform better than character-based models due to their unambiguousness in terms of pronunciation, transcript data was converted to this type of representation beforehand.

1) Text Modification

Since many punctuation symbols have very similar effects on emphasis in German, a subset was defined onto which all further symbols were mapped. This resulted in a subset consisting only of the characters [“.”, “,”, “?”, “!”], which significantly reduced of the vocabulary size.

Additionally, datasets based on LibriVox mostly consist of audio books of which the transcripts were written in the early 20th century and earlier, as German licensing rights require authors to have been deceased for at least 70 years, before copyright of their works expires. Transcripts of such ages were written according to obsolete orthographic standards, but the models to be trained were intended to be used in modern contexts. For this reason, a dictionary has been created semiautomatically (partly by crawling [27], a website providing common mappings between orthographic conventions, partly through manual identification of obsolete phrasing inside transcript sentences). Utilizing regular expressions, the outdated transcripts were adapted to currently applicable orthographic principles.

2) Phonemization

As no publicly available mapping tools or dictionaries seemed to be performing well enough for phonemization in German, a custom dictionary was created by crawling Wiktionary German [28], a website providing over 640,000 German word pairs with notations based on character as well as the International Phonetic Alphabet (IPA) including nouns in multiple grammatical cases and verbs in multiple tenses.

To convert composites which are not exactly contained within the phoneme dictionary into phoneme notation, a bidirectional search algorithm was implemented, which splits words into substrings if no exact match is found. The longest substrings found are individually converted to phoneme symbols and merged back together afterwards.

Since compounds and nominalizations by using different suffixes are widely used in the German language, a major proportion of the vocabulary can be covered by this approach.

While this algorithm handles borderline cases, names and words from other languages rather poorly, most German words as well as composites can be mapped to their respective phoneme representation quite efficiently. To reduce suboptimal mappings to a minimum, a large fraction of unknown words contained in the selected training datasets was added manually to the phoneme dictionary.

IV. MODEL TRAINING

The following subsections present and justify the final selection of model architectures for both stages of a full TTS system and describe all conducted training workflows on a detailed level. Both acoustic models and vocoders were trained independently.

A. Model Selection

Since a wide range of architectures exists for both acoustic models and vocoders, several test trainings were conducted to determine a viable composition.

Tacotron 2 and TransformerTTS were considered as acoustic models due to their excellent evaluations in literature as well as their inclusion into the ESPnet [29] framework, a toolkit for speech processing, offering simple mechanisms for building TTS training pipelines. First trainings showed that stop token prediction clearly performed better with Tacotron 2 than TransformerTTS, thus the final choice was made in favor of this architecture. AlignTTS was considered as well, but preexisting implementations were badly documented and training with reasonable effort was unfeasible.

For the vocoder stage, it was intended to test several architectures in sequence. However, Multi-Band MelGAN, as first architecture to be evaluated, already achieved subjectively satisfactory results in initial tests and was selected as the vocoder architecture for subsequent trainings. It was refrained from testing other vocoders, since subjectively, the quality of the acoustic model had a larger impact on overall output quality.

B. Tacotron 2

To optimize the training process, minor adjustments were made to the default hyperparameter configuration before the training process. In addition, the most suitable decoder configuration at inference time was determined through manual evaluation.

1) Training

The specific model architecture and training configuration for Tacotron 2 were derived from the existing recipe for LJSpeech incorporated in the ESPnet framework and adapted to fit the available hardware in terms of batch size (or number of batch bins, as implemented in ESPnet). This recipe differs from the original implementation of Tacotron 2 in the usage of guided attention loss. While training with datasets based on a sampling rate of 16kHz resulted in fast loss convergence, models trained on 22.05kHz audio data quickly reached a stage of oscillating loss. This was remedied by the use of AMSGrad [30]. All other parameters were maintained. In order to utilize ESPnet, the datasets used were converted into the Kaldi [31] format.

All models were trained for 200 epochs. Training processes took ~4 days each for 22.05kHz datasets and ~2 days each for 16kHz datasets respectively on two Nvidia Titan RTX GPUs.

2) Inferencing

The decoder configuration can be dynamically adjusted at inference time. In order to find the best possible configuration for all speakers, several suitable values were defined for each adjustable parameter and output audio was generated for each combination of parameters. Any of the variables may cause word repetition or deletion errors, if misconfigured.

The following parameters were determined:

- Minimum Length Ratio: 0.08
- Maximum Length Ratio: 10
- Backwards Attention Window: 2
- Forwards Attention Window: 3
- Stop-Token-Threshold: 0.1

While optimal values varied slightly between all speakers, the specified configuration generally yielded good results. This rendered the following model evaluations independent of speaker-specific decoder configurations.

C. Mutli-Band MelGAN

The implementation used was the publicly available version by Tomoki Hayashi [30] and the standard configuration was retained. Each model was trained according to this for 800,000 steps. Training took ~3 days per model using the same hardware as for the Tacotron 2 models.

For the speaker Hokuspokus no separate vocoder with the clean subset was trained, instead the vocoder from the full dataset was reused.

V. Empirical Analysis

The trained model compositions were evaluated through a survey, collecting MOS values for original speakers, full two-level inferences, and inferences of vocoders based on algorithmically generated mel spectrograms of original recordings. Additionally, the survey included further questions regarding the "best" fully synthetic voice, according to individual ratings of the respondents. Furthermore, demographic parameters, as well as audio output devices used during the survey were queried.

A. Questionnaire Design

The core components and structure of the survey are described in more detail in the following subsections.

1) MOS

Each respondent could listen to three audio files per voice, which were to be rated qualitatively on a scale of 1 to 9 without further instructions. No text labels for the individual numbers were provided on purpose, it was merely indicated that 9 meant very good and 1 very bad quality. Fully synthetic voices (acoustic model + vocoder), ground truths of all voices as well as vocoder-only inferences derived from mel spectrograms of ground truth data were evaluated in order to gain insights into the general performance of the model combinations as well as the sole influence of the trained vocoders on speech quality. The judgements of the mean opinion scores thus included 16 different voices and 48 audio recordings with 5-8 seconds length per recording.

During the evaluation, the rating scale was rescaled to the range 1 to 5 (in 0.5 increments) to enable direct comparison to the MOS values of other publications.

To avoid a bias regarding the order of the heard speakers, the sequence in which respondents were to rate them was randomized.

2) Detailed "Best" Speaker

After all MOS values had been filled in, the best-rated, fully synthetic voice was automatically determined, its corresponding recordings were played again, and more in-depth questions were asked regarding the characteristics of this voice.

- Did you notice any anomalies in pronunciation you found annoying? (*Very many* to *None*) (Q1)
- How would you describe the effort needed to understand the message? (*Nothing understood* to *Everything understood*) (Q2)
- How did you perceive the pace of speech? (*Too slow* to *Too fast*) (Q3)
- How did you perceive the naturality of the voice? (*Very unnatural* to *Very natural*) (Q4)
- Did you find certain words difficult to understand? (*Very many* to *None*) (Q5)
- How would you describe the voice? (*Very unpleasant* to *Very pleasant*) (Q6)
- Would you find it easy or difficult to listen to this speaker for an extended period of time? (*Very easy* to *Very difficult*) (Q7)

These questions were intended to provide insight into which aspects of the synthetic voices were subjectively perceived as suboptimal. The selection of questions was based on [32]. Posterior characters represent references to the questions in TABLE **V**.

3) Demographic Data

To derive further conclusions from previously collected scores, participants were additionally asked regarding their native language and age. As described in CrowdMOS [33], the audio device used while answering the survey was also asked for.

B. MOS Results

The survey was conducted over the internet. Invitations were sent to students from the University of Applied Sciences Hof, the research institute employees, as well as to a network of company partners. It was also circulated on the internet via Twitter and Linked.in.

A total of 193 participants was recorded of which 101 finished the survey. 94 of this subset were native German speakers. Answers and ratings of those were used for further analysis. Around half of the leftover respondents used a smartphone or PC with built-in speakers. 34 were using headphones, 11 dedicated loudspeakers. The age of participants was 30.1 years on average with a median of 26 and a range from 18 to 74 years.

TABLE **III** summarizes the results. Synth represents the MOS for the synthetic voice, created using both the trained acoustic model and vocoder. Vocoder represents the MOS for the synthetic voice that was generated based on the mel spectrograms derived from the ground truth. GT represents the MOS for the human speaker used as training data. Δ GT is the difference between the MOS of the ground truth and the MOS of the synthetic voice. TABLE **IV** puts the results and training datasets in relation to each other.

C. *Speaker-Specific Analysis*

The more detailed, speaker-specific analysis shown in TABLE **V** presents an overview of the advanced evaluation, including certain characteristics of speech, which primarily revealed a persistent deficit of naturalness in the voices, where no synthetic voice reached an average score over 4.0. This is supported by comparable scores for anomalies in pronunciation and how pleasant the voice is perceived. Comprehensibility of individual words was rated slightly better. Pace of speech and effort required to understand the message of utterances were rated very positively. Ultimately, scores for the difficulty of listening to a speaker over an extended period of time were consistently mediocre. Bernd Ungerer especially stood out regarding naturalness of the synthetic voice, whereas there was no large difference to other voices regarding anomalies in pronunciation and ease of understanding compared to Thorsten. The pace of speech was also similar.

VI. Discussion

The empirical survey affirmed the preexisting subjective impression that the fully synthetic TTS system trained on data from the speaker Bernd Ungerer produced the best results among all evaluated model compositions. However, the overall scores were lower than expected. This is partly due to a large variation in answers with participants voting 2.3 on average for all 16 voices and others voting 4.6 (avg: 3.55, median: 3.62). With 94 qualified answers, the empirical survey is much larger than the ones in other TTS papers that frequently use less than two dozen participants.

Interestingly, the speaker Thorsten Müller achieved best results for vocoder only and a similar distance between synthetic voice and ground truth as Bernd Ungerer, despite having only a quarter of the training data. This indicates that data quality is at least equally important, if not more important than total size of the dataset. The same conclusion can be drawn from the results of Hokuspokus clean and full. Although the clean subset contains only 27 hours of voice data, the MOS results are slightly better than those of networks trained on the full 43 hours of data available. Which amount of (qualitatively high) training data would actually be needed for a well performing acoustic model remains to be determined. Matsubara et al. [34] found that as few as one hour of training data is sufficient for achieving MOS values of 3.8 with LPCnet and 9 hours for MOS values of 4.06 with WaveNet, with a ground truth of only 4.18. However, this could not be reproduced using Tacotron 2 and Multi-Band MelGAN, which may be caused by the chosen model composition. Stop token prediction proved problematic, which resulted in additional babbling sounds as part of the generated audio files. This mainly occurred with models trained on less than 20 hours of audio-transcription pairs.

The ground truth values of 4.25 and 4.27 for speakers Bernd Ungerer and Hokuspokus (both full and clean) are similar to the values reported in literature for English language, e.g., 4.27 for FastSpeech 2 [35] and 4.31 for TalkNet [36]. However, they are significantly worse than the 4.58 reported in the Tacotron 2 paper [1] or 4.55 for Flow-TTS [11]. This indicates that there is still potential for improvement since neither Bernd Ungerer nor Hokuspokus are professional speakers. Accordingly, the recordings were not professionally produced and processed, which in consequence lead to inconsistent narration styles and noise. A delta of 0.5 between ground truth MOS and synthetic voice (Bernd Ungerer, Thorsten Müller) is only topped by very few of the well-known English TTS results published. It can therefore be concluded that the chosen model architectures can generally be equally well trained on datasets in German as in the English language (or Chinese for Multi-Band MelGAN).

Vocoder MOS values are significantly lower than expected for all speakers except Thorsten Müller. A delta of 0.22 for Thorsten Müller is among the best in published English results. However, for Bernd Ungerer (0.50) and Hokuspokus (0.66), values are worse than the average published in English publications concerning TTS, which is around 0.35. For Multi-Band MelGAN, the published results are 4.22, which is 0.36 worse than the ground truth on the MOS scale. However, these results were gathered in Chinese. Switching the vocoder should be investigated for future experiments.

Differently than suggested in literature [37], the female voices are not judged better than the male voices, but worse. This is especially unexpected for the direct comparison of Hokuspokus clean with Thorsten Müller. Hokuspokus has a better GT score and slightly more training data in the clean dataset (27h vs. 23h). Therefore, a better MOS value for Hokuspokus than for Thorsten was expected. There are two major differences between the datasets. The Thorsten neutral set consists of one (short) sentence per audio sample having an average duration of 3.3s with a maximum of 12s and only few audio files with more than 5s (see Figure **1**), whereas audio-transcripts from Hokuspokus (and other sets from the HUI audio corpus) were split based on duration with a minimum length of 5s and an average of 9s with some audio files at over 20s, regardless of sentence cohesion.

Utterances in the Thorsten neutral dataset are continuously very clearly emphasized as it was specifically generated for the creation of TTS systems, while recordings by Hokuspokus do not contain any special emphasis, sounding generally more natural (which possibly led to comparably higher MOS values for ground truth). However, this aspect seems to render the Hokuspokus datasets less suitable for speech synthesis applications. Additionally, the average silence loudness in dB is slightly lower in the Thorsten neutral dataset (-58.3 dB) compared to Hokuspokus clean (-56.6 dB, see Figure 2), indicating less noise. It would be interesting to see, whether a further cleansing of Thorsten speech samples yield better training results. Due to the generally low amount of training data contained in the Thorsten neutral dataset, no further investigation was conducted.

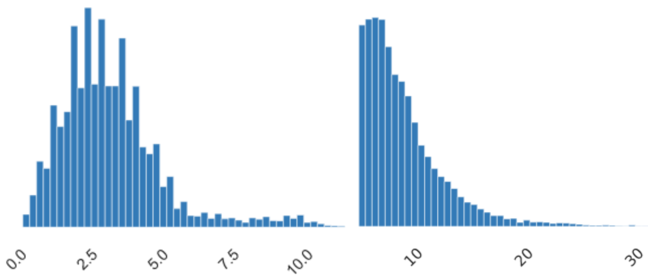

Figure 1. Thorsten (l) and Hokuspokus (r) length of audio in seconds.

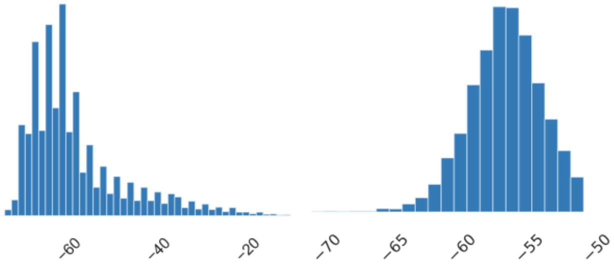

Figure 2. Thorsten (l) and Hokuspokus (r) min. silence in dB.

It is also surprising, that the speaker Karlsson achieved a comparably high MOS for the ground truth despite being based on a sampling rate of 16 kHz. Also, the vocoder MOS is among the best with 3.76, whereas the fully synthetic voice merely achieved 2.96 (-0.8 compared to the vocoder).

Moreover, it is remarkable that the loss in MOS from GT to vocoder and full synthetic voice is split relatively equally for Thorsten Müller and Hokuspokus, whereas there is nearly no loss for the acoustic model for Bernd. In contrast to that, Karlsson and Eva have most of the loss in acoustic model and a much smaller one for the vocoder. Looking at published results for LJspeech, examples of both described discrepancies can be found. For an equal split, there is AlignTTS, FlowTTS und TalkNet with WaveGlow vocoder, as well as TalkNet 2 with Hifi-GAN vocoder. A larger loss for the vocoder can be observed for Glow-TTS and Fastspeech with WaveGlow vocoder, as well as Reinforce-Aligner and Diff-TTS with Hifi-GAN vocoder. Finally, EFTS-CNN with Hifi-GAN has a higher loss in the acoustic model than the vocoder. Therefore, it could be a matter of tuning the hyperparameters for the training process that makes a difference, but it could also be characteristics of the dataset in this case. It is assumed that the acoustic model benefits more from large amounts of training data, whereas the vocoder benefits more from a high audio quality.

Furthermore, speaker-specific analysis confirmed that basic conditions for natural speech, such as pace and correct as well as clear pronunciation of individual words, are generally met. However, fully synthetic outputs still contain too many irregularities, which reduces the acceptance of users to listen over longer periods of time. Additionally, none of the recordings contained in the training datasets were made by a professional speaker, which is reflected in the mediocre scores on how pleasant the different voices were perceived.

VII. Limitations

Audio files, which were used for the empirical analysis were specifically chosen to be comparable across all speakers as well as comparable with the ground truth. Although sentences that proved to be difficult during the training process were included, they are still somehow cherry-picked. When generating speech from arbitrary texts from news websites, some problems with the synthesized voices were encountered that are not reflected in the test audio. Negative examples can be found on the webpage, presenting results ([38]).

Although these cases are seldom, the quality of the generated speech output still needs to be double-checked, since Tacotron 2 performs in a non-deterministic way, which is intended in order to vary stylistic attributes in output mel spectrograms. However, this feature sometimes leads to very bad output quality.

Additionally, the choice of vocoder and acoustic model are somewhat arbitrary. Although there was a systematic analysis of available models, no detailed evaluation with multiple candidates was performed. Instead, the first models subjectively producing good results were used for the empirical study. Finally, the vocoder should have been trained with the Hokuspokus clean subset as well, instead of reusing the one from the full subset in order to explore the full potential of data cleansing.

VIII. Conclusion and Outlook

In this work, the training processes of several deep neural networks for speech synthesis in the German language was reported along with an evaluation based on the MOS. A MOS of 3.74 was achieved for the best rated model (using the speaker Bernd Ungerer), which is comparable to recently published results for speech synthesis systems in English like 3.79 for FastSpeech 2 [35] or 3.66 for Flowtron [18].

However, they are far away from the best published results like 4.53 for Tactron 2 with Wavenet [1] or 4.19 for Flow-TTS with WaveGlow [11]. On the other hand, Tacotron 2 also achieves only 3.52 on the MOS scale in the Flow-TTS paper. To the best of the authors' knowledge, results are the best published MOS results for German TTS and can serve as a benchmark for future publications. In the years before neural TTS systems, MaryTTS has been a well-known option for German [39] and multi-lingual speech synthesis [40]. However, even in explicit quality analysis [41], no MOS values are reported.

In addition, deeper insights were gained regarding distinct aspects of different synthetic voices, which suggest actions regarding further optimization of future models. At dataset level, alignment of audio transcript pairs, recording quality and its homogeneity, as well as prosody can be improved. Regarding the definition of hyperparameters, values were set based on comparisons. A thorough hyperparameter search could lead to better results. In addition, the phoneme dictionary needs to be extended to include a larger number of terms in order to cover as many words as possible.

All compared models and respective recipes for ESPnet are released for public use.

For further research, it is intended to continue experimenting with internal voice datasets of higher quality but smaller size, as well as different network architectures. Especially for the vocoder, a broader range of alternatives to Multi-Band MelGAN will be considered, including Hifi-GAN [20], WaveGrad [21] and Wave RNN [22], which all have published results well over 4.3 MOS in English language and differences to ground truth below 0.1.

Additionally, it needs to be investigated which aspects of the training data differentiate a very good from an average dataset. A few aspects like good recording conditions and trained speaker are well known. However, there is little information regarding speaking style, choice of sentences and words, diversity of the vocabulary, etc. Those aspects are expected to influence dataset quality. Moreover, the preexisting processing pipeline for the generation of datasets from [25] will be altered to shorten the minimum and maximum duration of audio snippets contained in training data to a scope 2s minimum, 6s mean and 15s maximum.

Curriculum learning [42] represents another promising method, which would be worth investigating in the context of TTS. It is dangerous to draw conclusions from humans to DNNs. Despite some similarities, DNNs still work different from human brains. Nevertheless, human children usually learn to speak short utterances first, as opposed to words like "Frühsommer-Meningoenzephalitis" (FSME), a complex German word from the medical domain, which is part of an internal test dataset. Therefore, it could be also helpful to run trainings of model architectures with audio-transcription pairs of short sentences or even single words and gradually increase the length of labeled audio files. There is already evidence that this method increases robustness of TTS models for longer input texts during inference [43]. It could potentially also improve loss convergence during training as well as output speech quality.

The findings presented in this work will be incorporated into the development of an independent smart speaker, whereby the performance of TTS systems on edge devices, primarily resource requirements and RTF, will be a major challenge.

References

[1] 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.

[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," *arXiv preprint arXiv:2005.05106*, 2020.

[3] S. Kayte, M. Mundada, and J. Gujrathi, "Hidden Markov model based speech synthesis: A review," *International Journal of Computer Applications*, vol. 130, no. 3, pp. 35–39, 2015.

[4] Y. Wang *et al.*, "Tacotron: Towards End-to-End Speech Synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[5] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[6] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 01, pp. 6706–6713.

[7] A. Vaswani *et al.*, "Attention Is All You Need," *arXiv:1706.03762 [cs]*, Dec. 2017, Accessed: Sep. 01, 2021. [Online]. Available: http://arxiv.org/abs/1706.03762

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Neural Speech Synthesis with Transformer Network," 2019.

[10] Y. Ren *et al.*, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*, 2019, pp. 3171–3180.

[11] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7209–7213.

[12] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," *arXiv preprint arXiv:2005.11129*, 2020.

[13] I. Elias *et al.*, "Parallel Tacotron: Non-Autoregressive and Controllable TTS," *arXiv preprint arXiv:2010.11439*, 2020.

[14] A. Łańcucki, "FastPitch: Parallel Text-to-speech with Pitch Prediction," *arXiv preprint arXiv:2006.06873*, 2020.

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

[16] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.

[17] Z. Zeng, J. Wang, N. Cheng, T. Xia, and J. Xiao, "Aligntts: Efficient feed-forward text-to-speech system without explicit alignment," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6714–6718.

[18] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis," *arXiv preprint arXiv:2005.05957*, 2020.

[19] "The LJ Speech Dataset." https://keithito.com/LJ-Speech-Dataset (accessed Sep. 01, 2021).

[20] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *Advances in Neural Information Processing Systems*, vol. 33, n. pag., 2020.

[21] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," *arXiv preprint arXiv:2009.00713*, 2020.

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

[23] "LibriVox | free public domain audiobooks." https://librivox.org/ (accessed Sep. 01, 2021).

[24] T. Müller, "Thorsten Open German Voice Dataset". https://github.com/thorstenMueller/deep-learning-german-tts (accessed Sep. 01, 2021).

[25] P. Puchtler, J. Wirth, and R. Peinl, "HUI-Audio-Corpus-German: A high quality TTS dataset," Berlin, Germany, Sep. 2021.

[26] "Projekt Gutenberg". https://www.projekt-gutenberg.org/ (accessed Sep. 01, 2021).

[27] J. von Heyl, "korrekturen.de - Portal für Rechtschreibung". https://www.korrekturen.de/ (accessed Sep. 01, 2021).

[28] "Wiktionary, das freie Wörterbuch". https://de.wiktionary.org/wiki/Wiktionary:Hauptseite (accessed Sep. 01, 2021).

[29] T. Hayashi *et al.*, "Espnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 7654–7658. doi: 10.1109/ICASSP40776.2020.9053512.

[30] S. J. Reddi, S. Kale, and S. Kumar, "On the Convergence of Adam and Beyond," *arXiv:1904.09237 [cs, math, stat]*, Apr. 2019, Accessed: Sep. 01, 2021. [Online]. Available: http://arxiv.org/abs/1904.09237

[31] D. Povey *et al.*, "The Kaldi speech recognition toolkit," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Jan. 2011.

[32] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Computer Speech & Language*, vol. 19, no. 1, pp. 55–83, 2005, doi: https://doi.org/10.1016/j.csl.2003.12.001.

[33] F. Protasio Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "CROWDMOS: An Approach for Crowdsourcing Mean Opinion Score Studies," May 2011, Accessed: Sep. 01, 2021, ICASSP. [Online]. Available: https://www.microsoft.com/en-us/research/publication/crowdmos-an-approach-for-crowdsourcing-mean-opinion-score-studies/

[34] K. Matsubara *et al.*, "Investigation of training data size for real-time neural vocoders on CPUs," *Acoustical Science and Technology*, vol. 42, no. 1, pp. 65–68, 2021.

[35] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech," *arXiv preprint arXiv:2006.04558*, 2020.

[36] S. Beliaev, Y. Rebryk, and B. Ginsburg, "TalkNet: Fully-Convolutional Non-Autoregressive Speech Synthesis Model," *arXiv preprint arXiv:2005.05514*, 2020.

[37] J. Cambre and C. Kulkarni, "One voice fits all? Social implications and research challenges of designing voices for smart devices," *Proceedings of the ACM on human-computer interaction*, vol. 3, no. CSCW, pp. 1–19, 2019.

[38] J. Wirth, "iisys Audio Samples for German Speech Synthesis Tacotron 2 + MultiBand MelGAN". http://narvi.sysint.iisys.de/projects/tts/results (accessed Sep. 01, 2021).

[39] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[40] M. Schröder, M. Charfuelan, S. Pammi, and I. Steiner, "Open source voice creation toolkit for the MARY TTS Platform," presented at the Twelfth annual conference of the international speech communication association, 2011.

[41] F. Hinterleitner, C. Norrenbrock, S. Möller, and U. Heute, "What makes this voice sound so bad? A multidimensional analysis of state-of-the-art text-to-speech systems," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 240–245.

[42] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[43] S.-W. Hwang and J.-H. Chang, "Document-Level Neural TTS Using Curriculum Learning and Attention Masking," *IEEE Access*, vol. 9, pp. 8954–8960, 2021.

TABLE II. MOS OVERVIEW OF COMPARABLE TTS SYSTEMS.

Model	Vocoder	GT	Vocoder	Synth	GT-MOS	GT-voc	Voc-synth
Fastspeech	WaveGlow	4.41	4.00	3.84	0.57	0.41	0.16
AlignTTS	WaveGlow	4.53	4.28	4.05	0.48	0.25	0.23
Glow-TTS	WaveGlow	4.54	4.19	4.01	0.53	0.35	0.18
Flow-TTS	WaveGlow	4.55	4.35	4.19	0.36	0.20	0.16
TalkNet	WaveGlow	4.31	4.04	3.74	0.57	0.27	0.3
TalkNet 2	Hifi-GAN	4.32	4.2	4.08	0.24	0.12	0.12

TABLE III. MOS COMPARISON OF ALL TRAINED SPEAKERS.

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

TABLE IV. OVERVIEW OF DATASETS USED FOR MODEL TRAINING AND CORRESPONDING MOS EVALUATIONS.

Speaker	GT	Δ GT-synth	Δ GT-Vocoder	Δ Vocoder-synth	Amount of data (hours)	Training Loss (Acoustic Model)	Sampling Rate
Bernd Ungerer	4.25	0.51	0.50	0.01	97	0.52	22.05 kHz
Thorsten Müller	3.99	0.50	0.22	0.28	23	0.48	22.05 kHz
Hokuspokus Clean	4.27	1.29	0.66	0.62	43	0.44	22.05 kHz
Hokuspokus Full	4.27	1.39	0.66	0.72	27	0.46	22.05 kHz
Karlsson	4.14	1.18	0.38	0.80	40	0.43	16 kHz
Eva K.	3.72	1.60	0.39	1.20	29	0.56	16 kHz

TABLE V. SPEAKER-SPECIFIC ANALYSIS (OPTIMAL SCORES IN BRACKETS).

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

- Did you notice any anomalies in pronunciation you found annoying? (*Very many* to *None*) (Q1)
- How would you describe the effort needed to understand the message? (*Nothing understood* to *Everything understood*) (Q2)
- How did you perceive the pace of speech? (*Too slow* to *Too fast*) (Q3)
- How did you perceive the naturality of the voice? (*Very unnatural* to *Very natural*) (Q4)
- Did you find certain words difficult to understand? (*Very many* to *None*) (Q5)
- How would you describe the voice? (*Very unpleasant* to *very pleasant*) (Q6)
- Would you find it easy or difficult to listen to this speaker for an extended period of time? (*Very easy* to *Very difficult*) (Q7)