18th Blizzard Challenge Workshop
29 August 2023, Grenoble, France

The AudioLabs System for the Blizzard Challenge 2023

Frank Zalkow[1], *Paolo Sani*[1], *Michael Fast*[1], *Judith Bauer*[1], *Mohammad Joshaghani*[1], *Kishor Kayyar Lakshminarayana*[1,2], *Emanuël A. P. Habets*[1,2], *Christian Dittmar*[1]

[1]Fraunhofer IIS, Erlangen, Germany [2]International Audio Laboratories Erlangen†, Germany

frank.zalkow@iis.fraunhofer.de

Abstract

In this paper, we describe our contribution to the Blizzard Challenge 2023. This challenge has the goal of understanding and comparing research techniques in building corpus-based speech synthesizers on the same data. The 2023 edition of the challenge focuses on the French language and low-resource settings. Our text-to-speech (TTS) synthesis system consists of three main building blocks. First, a non-autoregressive acoustic model converts symbolic input sequences (phonemes) into mel-scaled speech spectrograms. Second, a post-processing model based on a generative adversarial network (GAN) enhances the predicted mel spectrograms. Third, the GAN-based neural vocoder StyleMelGAN converts the enhanced spectrogram into a time-domain speech waveform.

Index Terms: Blizzard Challenge, text-to-speech synthesis

1. Introduction

Text-to-speech synthesis (TTS) aims to produce synthetic speech waveforms from text inputs. In recent years, deep-learning-based methods have led to considerable improvements of the audio quality, naturalness, and intelligibility of synthetic speech in comparison to the previous generation of approaches (e.g., based on hidden Markov models [1]). A multitude of different architectures and configurations of such TTS systems has been proposed, ranging from multi-step approaches (such as the combination of an acoustic model and a neural vocoder [2]) to complete end-to-end systems [3]. The Blizzard Challenge, a long-established competition among research groups working on TTS, helps to compare and understand the differences between these approaches. In the challenge, a training corpus is released, which all participants use to train their TTS system. After the participants submit synthesized speech for a given set of test prompts, double-blind listening tests are conducted to evaluate the respective systems' quality. Through this procedure, a fair comparison between the different participating systems can be ensured.

This year's challenge focuses on the French language with two different tasks. The first task is to build a synthetic voice from a typical, well-resourced training corpus of a single female French speaker. The total duration of this dataset is more than 51 hours. The second task aims at building a synthetic voice of another female French speaker, but the training corpus is much smaller, covering about two hours. Having a smaller amount of training data makes the task more difficult. However, this task can still be considered to be of moderate difficulty compared to extremely low-resource settings, where only a few minutes of training data are available [4].

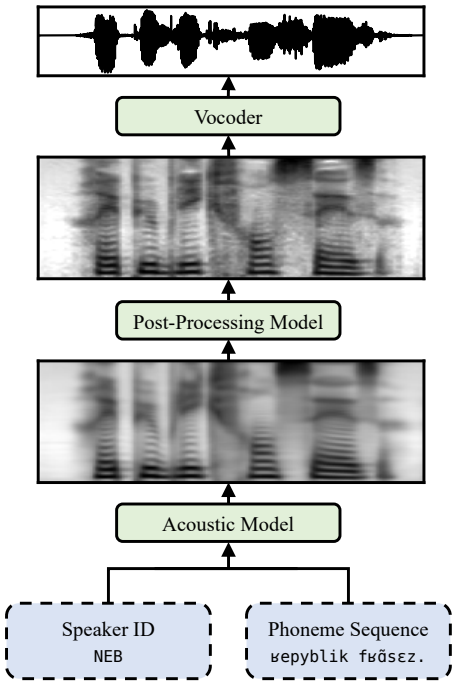

Figure 1: *Overview of our TTS system.*

In this paper, we describe the system that we used to participate in the challenge. Our TTS system consists of three neural networks: First, we use an acoustic model that predicts mel-scaled spectrograms from symbolic inputs in phoneme representation. The network architecture is inspired by Forward-Tacotron [5] and FastTacotron [6]. Second, we use a post-processing model based on a generative adversarial network (GAN) that enhances the naturalness of the spectral representations from the acoustic model [7]. The architecture for this module takes inspiration from the Pix2Pix system [8, 9]. Third, we use another GAN-based network, called StyleMelGAN [10], that acts as a neural vocoder to convert the enhanced mel spectrograms to a time-domain speech waveform.

The remainder of the paper is structured as follows. In Section 2, we explain our TTS system and its components. Then, we describe the used dataset, our text processing, and the extraction of prosody-related variation features from the speech recordings in Section 3. Next, in Section 4, we comment on the performance of our system in the challenge's evaluation. Finally, we close in Section 5 with a short conclusion.

†A joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

10.21437/Blizzard.2023-8

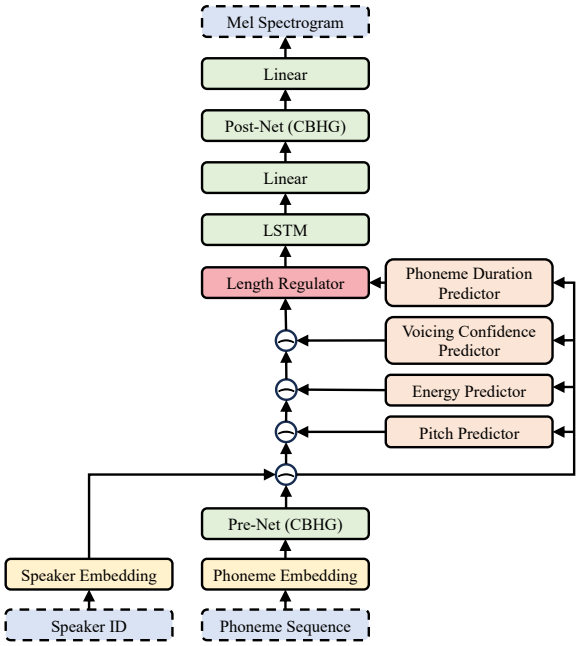

Figure 2: *Overview of our acoustic model. (⌢ refers to concatenation. CBHG refers to a network module originally introduced for Tacotron [11].)*

Layer	Output Size	Parameters
Speaker Embedding	$(N, 64)$	320
Phoneme Embedding	$(N, 256)$	15 104
Pre-Net (CBHG)	$(N, 512)$	13 580 288
Pitch Predictor	$(N, 1)$	3 760 641
Energy Predictor	$(N, 1)$	3 760 641
Voicing Confidence Predictor	$(N, 1)$	3 760 641
Phoneme Duration Predictor	$(N, 1)$	4 466 177
Length Regulator	$(M, 579)$	0
LSTM	$(M, 576)$	5 331 456
Linear	$(M, 80)$	92 240
Post-Net (CBHG)	$(M, 512)$	3 712 672
Linear	$(M, 80)$	40 960
		Σ: 38 521 140

Table 1: *Details of our acoustic model (N refers to the number of phoneme symbols in the input sequence and M refers to the number of spectral frames of the output).*

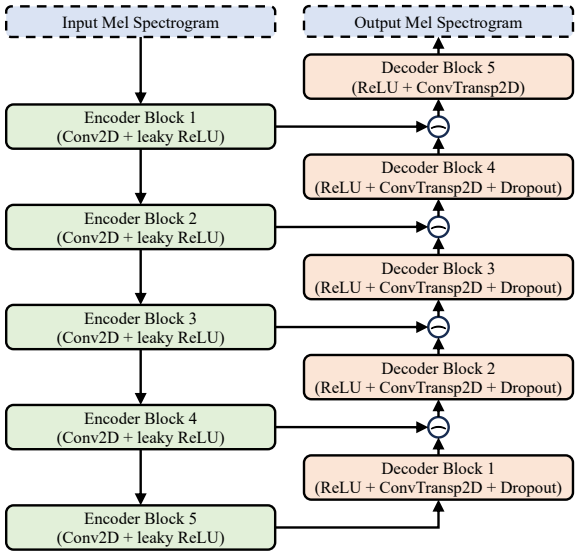

Figure 3: *Overview of our post-processing model (⌢ refers to concatenation).*

Layer	Output Size	Parameters
Encoder		
Block 1 (Conv2D)	$(80, M, 32)$	544
Block 2 (Conv2D)	$(40, M/2, 64)$	32 832
Block 3 (Conv2D)	$(40, M/2, 128)$	131 200
Block 4 (Conv2D)	$(20, M/4, 256)$	524 544
Block 5 (Conv2D)	$(20, M/4, 256)$	1 048 832
Decoder		
Block 1 (ConvTransp2D)	$(20, M/4, 256)$	1 048 832
Block 2 (ConvTransp2D)	$(40, M/2, 128)$	1 048 704
Block 3 (ConvTransp2D)	$(40, M/2, 64)$	262 208
Block 4 (ConvTransp2D)	$(80, M, 32)$	65 568
Block 5 (ConvTransp2D)	$(80, M, 1)$	1025
		Σ: 4 164 289

Table 2: *Details of our post-processing model (M refers to the number of spectral frames).*

2. TTS System

In this section, we explain the three parts of our TTS system in more detail. As already mentioned, it consists of an acoustic model (Section 2.1), a post-processing model (Section 2.2), and a neural vocoder (Section 2.3). Please note that the acoustic model architecture was never outlined in a publication before. In contrast, we use the post-processing model and neural vocoder similarly as described in the original publications [7, 10].

2.1. Acoustic Model

Our multi-speaker acoustic model [12] has a parallel neural architecture that uses text in phoneme representation following the international phonetic alphabet (IPA) and a speaker identifier [13] as input and predicts a mel-scaled spectrogram with logarithmic magnitude compression. The model is inspired by ForwardTacotron [5], which we extended by adding semantically meaningful variation predictors to model prosody, similar to FastTacotron [6]. Figure 2 shows an overview of our acoustic model, and Table 1 shows further details.

All variation predictors have the same architecture, inspired by the phoneme duration predictor from ForwardTacotron [5], which consists of a stack of three convolutional layers, a gated recurrent unit (GRU) [14], and a final linear layer. They output semantically meaningful prosody variation values and are trained using ground-truth values extracted from the training data (see Section 3.3).

2.2. GAN-Based Post-Processing Model

It is a well-known problem that the output of a parallel acoustic model often looks blurry, which is referred to as the "oversmoothing effect" [15]. Since it is common practice to train neural vocoders on ground-truth speech spectrograms, there is a mismatch between the input representation during training and inference, which may lead to audible artifacts in the final output of the TTS system. To compensate for this issue, we employ

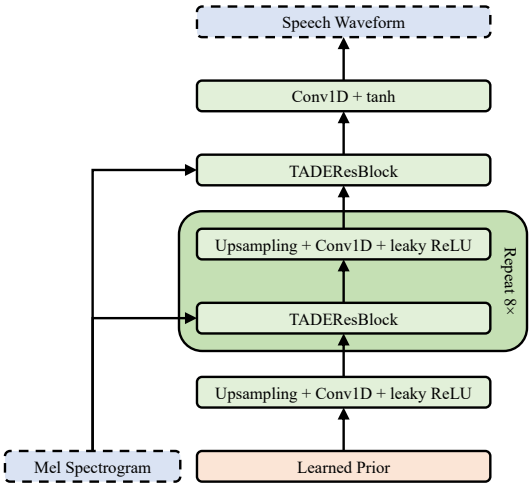

Figure 4: *Overview of the StyleMelGAN vocoder. TADEResBlock refers to TADE residual block as defined in [10].*

a GAN-based post-processing network [7], where the generator is based on the Pix2Pix architecture [8, 9]. As indicated in Figure 1, this model uses the output from the acoustic model as input and predicts an enhanced spectrogram with more natural fine-grained details compared to its input. Figure 3 shows an overview of the generator, and Table 2 shows further details.

2.3. Neural Vocoder

To convert the mel spectrogram into a speech signal, we use the neural vocoder StyleMelGAN [10]. Compared to the configuration described in the original publication, we use 128 channels throughout all TADE (temporal adaptive denormalization) residual blocks in the generator. Furthermore, we use a learned embedding (fixed during inference) instead of randomly sampled noise as prior. Figure 4 shows an overview of the StyleMelGAN generator, and Table 3 shows further details.

We only used the data described in Section 3.1 to train the vocoder. A larger and more diverse training corpus generally leads to better audio quality in the sense of a universal vocoder [16].

2.4. Technical Details

We use a consistent sampling rate of 22 050 Hz for the speech recordings. As a feature representation, we use logarithmically compressed mel-scaled spectrograms with 80 bands having center frequencies spread between 0 and 8 kHz, with 46.4 ms blocksize and 11.6 ms hop size.

We trained our acoustic model for 300k training iterations with a batch size of 64, our GAN-based post-processing model for 6.8 million training iterations with a batch size of 16, and our vocoder for about 1.5 million iterations with a batch size of 32. More specifically, the vocoder was trained for 941.2k iterations using only a multi-scale spectral reconstruction loss. Afterward, the GAN-based adversarial loss using an ensemble of four discriminators was used in addition to the reconstruction loss, as described in the original paper [10].

After training the acoustic model, we observed some unnatural prosody during inference, which we attributed to overfitting of the variation predictors. Therefore, we re-initialized the pitch, energy, voicing confidence, and phoneme duration predictors and retrained them while the rest of the acoustic model was fixed. While retraining, we stored the model weights and validation losses for each variation predictor separately. To ensure the reliability of the validation loss, we used an increased validation set of 2048 samples compared to the initial training, where only five samples were used. After retraining, we combined the model weights such that we used the weights for each of the variation predictors corresponding to the lowest respective validation loss. Table 4 shows the specific epochs and training iterations that correspond to the used weights.

Layer	Output Size	Parameters
Prior Embedding		
Learned Prior	$(1, 128)$	128
Conv1D	$(M, 128)$	147 584
Block 1		
TADEResBlock	$(M, 128)$	1 365 248
Conv1D	$(2M, 128)$	147 584
Block 2		
TADEResBlock	$(2M, 128)$	1 365 248
Conv1D	$(4M, 128)$	147 584
Block 3		
TADEResBlock	$(4M, 128)$	1 365 248
Conv1D	$(8M, 128)$	147 584
Block 4		
TADEResBlock	$(8M, 128)$	1 365 248
Conv1D	$(16M, 128)$	147 584
Block 5		
TADEResBlock	$(16M, 128)$	1 365 248
Conv1D	$(32M, 128)$	147 584
Block 6		
TADEResBlock	$(32M, 128)$	1 365 248
Conv1D	$(64M, 128)$	147 584
Block 7		
TADEResBlock	$(64M, 128)$	1 365 248
Conv1D	$(128M, 128)$	147 584
Block 8		
TADEResBlock	$(128M, 128)$	1 365 248
Conv1D	$(256M, 128)$	147 584
Final Block		
TADEResBlock	$(256M, 128)$	1 365 248
Conv1D	$(256M, 1)$	1153
		Σ: 13 616 769

Table 3: *Details of the StyleMelGAN vocoder (M refers to the number of spectral frames in the mel spectrogram).*

Variation Predictor	Epoch	Iteration
Pitch	29	63 684
Energy	13	28 548
Voicing Confidence	27	59 292
Phoneme Duration	2	4392

Table 4: *Epoch respective training iteration after which the weight was used for the different variation predictors.*

3. Data Preparation

3.1. Dataset

To train the components of our TTS system (explained in Section 2), we used a combination of the provided data from the Blizzard challenge and the publicly available SIWIS French dataset [17], which was created in the context of the SIWIS project [18]. We excluded some samples from our training set with non-regular speech, such as artistic speech, singing, or

Dataset	Processing	Speaker	Duration
Blizzard2023	none	NEB	51:29:13
Blizzard2023	denoised	NEB	51:28:52
Blizzard2023	none	AD	02:04:53
Blizzard2023	denoised	AD	02:04:52
SIWIS French	none	N/A	10:44:48
			Σ: 117:52:38

Table 5: *Data used to train our TTS system. NEB refers to Nadine Eckert-Boulet and AD refers to Aurélie Derbier.*

whispering. To identify these problematic samples in a semi-automatic fashion, we employed a speech–phoneme alignment system [19] based on the connectionist temporal classification (CTC) loss [20]. We trained the aligner and considered samples with a high loss and samples with sequences of short phoneme durations as candidates that need to be checked manually. Furthermore, we augmented the Blizzard2023 data by applying a commercially available denoiser from the Fraunhofer upHear product family[1]. We excluded a few samples where the denoising led to notable artifacts. Table 5 summarizes the used training data. When considering the original and the denoised version of a speaker as two different speakers, our combined dataset features five female French speakers of about 118 hours in total.

3.2. Text Processing

To convert text input (of both the training and test set) into a phoneme representation in IPA format, we used two sources: First, we used a custom pronunciation lexicon, combining a Wiktionary-based lexicon[2] and the annotations provided by the challenge. Second, if a word was not found in the lexicon, i.e., out-of-vocabulary, we used the open-source grapheme-to-phoneme (g2p) converter eSpeak[3].

3.3. Variation Data

To model the prosody of the speech, our acoustic model features semantically meaningful variation predictors that estimate pitch, energy, voicing confidence, and phoneme durations. These predictors were trained with variation data extracted from the recordings of the training set.

The phoneme durations were estimated with the aforementioned CTC-based speech–phoneme alignment system [19]. We first trained this aligner on the entire dataset and then fine-tuned it for each speaker. This fine-tuning led to improvements in the accuracy of the phoneme duration estimates. Figure 5 shows an example that illustrates this improvement. At the top is a visualization of a mel-scaled spectrogram of a speech recording, and below are visualizations for two different corresponding phoneme duration estimations. The upper phoneme durations are computed by the aligner before to the fine-tuning process and show an incorrect placement of the plosive phoneme /k/ into a silence region around second 1.0. The lower phoneme durations are computed by the aligner after the speaker-specific fine-tuning and correctly assign the space symbol to the silence region, followed by the short plosive.

[1] https://www.iis.fraunhofer.de/en/ff/amm/prod/upHear.html

[2] https://www.wiktionary.org

[3] https://espeak.sourceforge.net

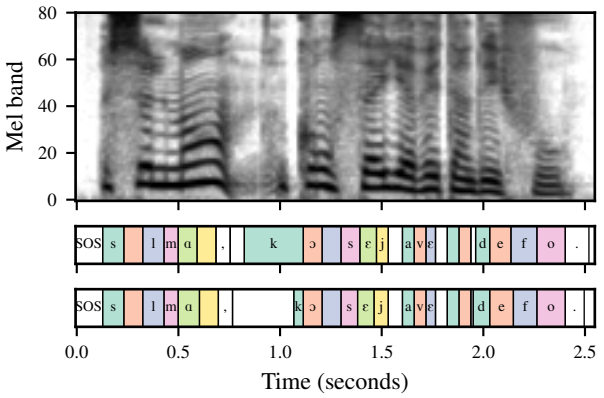

Figure 5: *Ground-truth mel spectrogram corresponding to the spoken prompt "Seulement, Conseil avait un défaut." (/sœlmɑ̃, kɔ̃sɛj avɛ d̃ɐ̃ defo./). Below are visualizations for two different phoneme duration estimations corresponding to the temporal axis of the spectrogram. In the upper and lower ones, the aligner without and with speaker-specific fine-tuning has been used, respectively. SOS refers to the beginning of a prompt. Note that the symbol for the end of the prompt (EOS) is not shown because we only use an excerpt of the prompt for visualization purposes.*

To estimate the pitch (i.e., fundamental frequency) trajectory of the recordings and the voicing confidence, we used CREPE [21]. Energy is just computed as the ℓ^2-norm of the magnitude values of each spectral frame. The pitch, energy, and voicing confidence values were then averaged across the frames corresponding to individual phonemes to obtain phoneme-level variation data.

4. Evaluation

4.1. Tasks

This year's Blizzard challenge features two different tasks. In both tasks, the aim is to create a model able to synthesize speech in the voice of the speaker from a particular training corpus. In the so-called "Hub" task, the speaker is NEB (see Table 5), where well-resourced training data is provided. In the so-called "Spoke" task, the speaker is AD, where low-resourced training data is provided.

We used the same multi-speaker TTS system trained on the speakers listed in Table 5 for both tasks of the challenge. Since the challenge's evaluation does not only focus on audio quality but also the similarity between the synthesized speech and the real recordings, we decided to imitate the acoustic conditions of the ground-truth recordings by using the speaker identifiers corresponding to the original data (and not the denoised versions) to generate the samples for our submission.

4.2. Listening Test

Eighteen and fourteen systems have been submitted for the Hub and Spoke task, respectively, where the system described in this paper is referred to by the letter "R." The quality of the TTS systems submitted to the challenge has been evaluated in listening tests.

As the first aspect, the listeners rated the quality of synthetic samples of the submitted TTS systems on a scale from

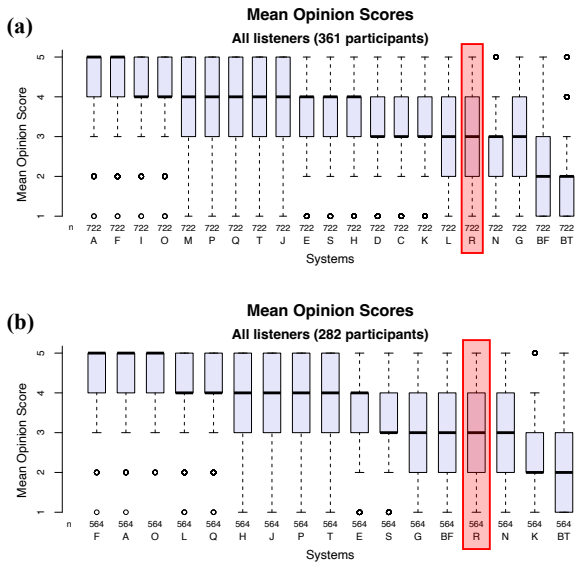

Figure 6: *Boxplots for distributions of MOS ratings for the quality within* ***(a)*** *the Hub and* ***(b)*** *Spoke task.*

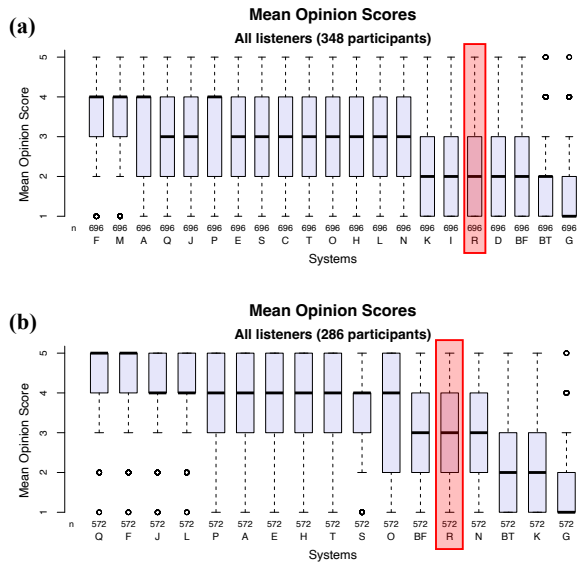

Figure 7: *Boxplots for distributions of MOS ratings for the speaker similarity within* ***(a)*** *the Hub and* ***(b)*** *Spoke task.*

1 ("very poor") to 5 ("excellent"), using a mean opinion score (MOS) test [22]. Figures 6a and 6b show boxplots for the resulting score distributions for the Hub and Spoke task, respectively. Our system R yielded a median score of 3 ("fair") in both tasks.

As a second aspect, the listeners were asked to rate the similarity of the synthetic voices compared to recordings of the real speaker. Again a MOS test was conducted, where the scale of ratings ranged from 1 ("completely different person") to 5 ("exactly the same person"). The results for the Hub task are shown in Figure 7a. Here, the MOS ratings are generally lower and never exceed a median of 4 for all submitted TTS systems. Our system yielded a median score of 2 ("probably a different person"). The results for the Spoke task are shown in Figure 7b. Here, the scores are higher for all systems, and our system yielded a median score of 3 ("similar").

Further aspects were also evaluated, such as the intelligibility of the synthesized speech samples. Furthermore, a MUSHRA test [23] of the top-performing TTS system has been conducted. We do not discuss these further evaluations here and refer to the organizer's paper about this year's challenge.

4.3. Discussion

Our system only yielded medium results in the challenge's evaluation and is not among the top-performing systems. We attribute this lack of performance to the relatively noisy training dataset. Moreover, we observed varying recording conditions and speaking styles in the training data. We usually rule this out by commissioning speech recordings to be performed under reproducible recording conditions. Our system is tuned to perform well on high-quality datasets, which we typically employ for production model development. It appears that the system can not cope very well with situations where the quality of the training data is lower.

To verify that the provided data has a high noise level, we employ a pre-trained a priori signal-to-noise ratio (SNR) estimator [24] using a multi-head attention network [25]. In particular, we use the DeepXi software package[4] and the pre-trained model "mhanet-1.1c." The average estimated a priori SNR $\hat{\xi}$ is 18.5 for speaker NEB and 13.1 dB for speaker AD. In contrast, a typical production dataset of ours yields an average $\hat{\xi}$ value of 31.2 dB.

Data augmentation with denoised samples as described in Section 3.1 did not yield the expected quality improvements. Further investigation is required to understand the reasons for that.

[4] https://github.com/anicolson/DeepXi

5. Conclusion

In this paper, we described our TTS system submitted to the Blizzard Challenge 2023, which focused on French data and low-resource settings. Our system performed sub-optimal in the challenge's evaluation, which we attribute to the fact that our system is tuned to work with high-quality datasets and the training set was noisy. In this sense, our participation was a valuable experience for our team and gave us further directions for improvement in case we want to enhance our system's robustness against noise.

The memory and run-time properties of the TTS systems are not considered in the challenge's evaluation. Our system is proven to be lightweight [12], i.e., it has low memory requirements and is much faster than real-time on both CPU (2.5×) and GPU (50.3×). Although such factors are not accounted for in the challenge's evaluation, they are crucial in practical applications.

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 Climate Action. In addition, this work was supported by the Free State of Bavaria in the DSAI project. The authors gratefully acknowledge the technical support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the FAU.

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 4779–4783.

[3] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *Proceedings of the International Conference on Learning Representations (ICLR)*, virtual, Austria, 2021.

[4] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T. Liu, "LRSpeech: Extremely low-resource speech synthesis and recognition," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, virtual, CA, USA, 2020, pp. 2802–2812.

[5] C. Schäfer, O. McCarthy, and contributors, "ForwardTacotron," https://github.com/as-ideas/ForwardTacotron, 2020.

[6] D. V. Sang and L. X. Thu, "FastTacotron: A fast, robust and controllable method for speech synthesis," in *Proceedings of the International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, Hanoi, Vietnam, 2021.

[7] P. Sani, J. Bauer, F. Zalkow, E. A. P. Habets, and C. Dittmar, "Improving the naturalness of synthesized spectograms for TTS using GAN-based post-processing," in *Proceedings of the ITG Conference on Speech Communication*, Aachen, Germany, 2023.

[8] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 5967–5976.

[9] P. Neekhara, C. Donahue, M. S. Puckette, S. Dubnov, and J. J. McAuley, "Expediting TTS synthesis with adversarial vocoding," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Graz, Austria, 2019, pp. 186–190.

[10] A. Mustafa, N. Pia, and G. Fuchs, "StyleMelGAN: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, Canada, 2021, pp. 6034–6038.

[11] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Stockholm, Sweden, 2017, pp. 4006–4010.

[12] P. Govalkar, A. Mustafa, N. Pia, J. Bauer, M. Yurt, Y. Özer, and C. Dittmar, "A lightweight neural TTS system for high-quality German speech synthesis," in *Proceedings of the ITG Conference on Speech Communication*, virtual, 2021, pp. 39–43.

[13] K. K. Lakshminarayana, C. Dittmar, N. Pia, and E. A. P. Habets, "Multi-speaker text-to-speech using ForwardTacotron with improved duration prediction," in *Proceedings of the ITG Conference on Speech Communication*, Aachen, Germany, 2023.

[14] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1724–1734.

[15] Y. Ren, X. Tan, T. Qin, Z. Zhao, and T. Liu, "Revisiting oversmoothness in text to speech," in *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, Dublin, Ireland, 2022, pp. 8197–8213.

[16] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," in *Proceedings of the International Conference on Machine Learning (ICML)*, Kigali, Rwanda, 2023.

[17] P.-E. Honnet, A. Lazaridis, P. N. Garner, and J. Yamagishi, "The SIWIS French speech synthesis database – design and recording of a high quality French database for speech," University of Edinburgh, School of Informatics, The Centre for Speech Technology Research, Tech. Rep., 2017.

[18] J. Goldman, P. Honnet, R. A. J. Clark, P. N. Garner, M. Ivanova, A. Lazaridis, H. Liang, T. Macedo, B. Pfister, M. S. Ribeiro, E. Wehrli, and J. Yamagishi, "The SIWIS database: A multilingual speech database with acted emphasis," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, San Francisco, CA, USA, 2016, pp. 1532–1535.

[19] F. Zalkow, P. Govalkar, M. Müller, E. A. P. Habets, and C. Dittmar, "Evaluating speech–phoneme alignment and its impact on neural text-to-speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

[20] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, Pittsburgh, Pennsylvania, USA, 2006, pp. 369–376.

[21] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 161–165.

[22] International Telecommunications Union, "Recommendation ITU-T P.800.1: Mean opinion score (MOS) terminology," Geneva, Switzerland, recommendation, 2016.

[23] ——, "Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," Geneva, Switzerland, recommendation, 2015.

[24] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, 2019.

[25] ——, "Masked multi-head self-attention for causal speech enhancement," *Speech Communication*, vol. 125, pp. 80–96, 2020.