The CPQD-Unicamp system for Blizzard Challenge 2021

Mário Uliani Neto[1], *Flávio O. Simões*[1], *Fernando O. Runstein*[1], *Edson J. Nagle*[1], *Bianca Dal Bó*[1], *Lucas H. Ueda*[1 2], *Paula D. P. Costa*[2]

[1]CPQD, Campinas, Brazil
[2]Department of Computer Engineering and Automation, School of Electrical and Computer Engineering, University of Campinas (UNICAMP), Campinas, Brazil

{uliani,simoes,runstein,nagle,bdalbo}@cpqd.com.br, l156368@dac.unicamp.br, paulad@unicamp.br

Abstract

This paper presents the CPQD-UNICAMP text-to-speech system for Blizzard Challenge 2021. The system consists of a bilingual linguistic front-end, an acoustic model based on Tacotron2 and a Parallel Wavegan neural vocoder. A multi-speaker Brazilian Portuguese dataset was added to the Blizzard 2021 dataset in order to train a bilingual acoustic model. The system was later fine-tuned with the target speaker data. Sentences were classified according to the punctuation type and a specialized model was trained for each category to better model the intonation pattern of non-declarative sentences. The Blizzard Challenge evaluation for the hub task shows that the proposed strategy achieved high naturalness, intelligibility and similarity results.

Index Terms: Blizzard Challenge 2021, Speech Synthesis, Tacotron2, Parallel Wavegan

1. Introduction

The Blizzard Challenge [1] has been organized annually since 2005, to better understand and compare research techniques in building corpus-based speech synthesizers on the same data.

In each edition, a speech dataset is provided to the challenge participants, who can use this data to build their systems according to the challenge rules.

This year's challenge [2] focused on the European Spanish language. Five hours of speech data from a female native speaker were provided. The challenge consisted of taking the released speech data, building a synthetic voice and synthesizing a prescribed set of test sentences. Two tasks were proposed: a *hub task*, in which participants should build a system to synthesise texts containing only Spanish words, and a *spoken* task, in which test sentences contained also a small number of English words.

Teams could submit synthetic speech for either or both tasks. In the case of this work, we focused exclusively on the hub task.

Techniques based on deep learning have been prevalent in the latest Blizzard evaluations [3], [4], consistently obtaining better results in terms of naturalness, intelligibility and similarity with the target voice. Most state-of-the-art systems today consist of a seq2seq acoustic model followed by a neural vocoder. The strategy presented in this work follows the same approach.

Generating a robust acoustic model requires a large amount of training data, which is often unfeasible when dealing with low resourced languages. In the case of the Blizzard 2021 challenge, the speech corpus consists of around five hours of recorded data, which, from our experience, we consider a reasonable amount of data, but still insufficient to build a high quality model. The use of additional data from different speakers in the same language is a strategy already explored in some works such as [5], [6] and [7] to overcome the lack of training data from a single speaker. However, for many languages of interest, the amount of data necessary to train a multi-speaker model is still limited or does not exist at all. Hence, other works tried to explore the use of multi-language multi-speaker modeling to achieve a better performance in the low-resource language scenario. This approach has proven to be successful: [8], [9] and [10], for instance, showed that mixing data from different languages can not only help to model low-resource languages but also improve their quality.

The approach used in this work consists of augmenting the provided material with non-Spanish data. First, a model trained exclusively with the challenge data was created to be used as a baseline. Next, experiments using both Brazilian Portuguese (BP) and American English (AE) as additional data were performed.

Considering that both European Spanish (ES) and BP are Ibero-Romance languages and share the same ancestor, while AE is a Germanic language, the first two share many similarities in lexicon, grammar and pronunciation. For example, the set of vowels and vowel clusters in ES can be considered practically a subset of BP, with very similar pronunciation. In AE, on the other hand, vowels are often pronounced as diphthongs and the pronunciation and intonations used are significantly different from both ES and BP. These similarities and differences between ES and the other two languages can be seen when analyzing the proprietary phoneme set used in the linguistic front-end of the proposed system, in which the set of phonemes in ES is much better represented within the set of BP phonemes than in the AE phonemes.

We hypothesize that the similarity between the target language and the language used for augmentation has an influence on the final result, especially with regard to the acoustic model. Such hypothesis was confirmed by our results, which showed a clear preference for the system that combined BP and ES in the training data.

Using both Blizzard and external data to train a bilingual multi-speaker acoustic model, which was later fine-tuned with the target speaker's data, proved to be very effective. However, there was still room to improve the quality of non-declarative sentences. Hence, we also propose dividing non-declarative sentences into four categories (three interrogative and one exclamatory) and creating a specialized acoustic model for each category.

Regarding the neural vocoder, a universal model was initially trained with 35 hours of multi-speaker multi-language speech data, and several fine-tuning approaches were tested to adapt this model towards the target speaker's voice.

The rest of this paper is organized as follows: Section 2 describes the data used in our experiments and outlines the building blocks of the proposed system. Section 3 describes the experimental setup, including data preparation and training strategies used to create the acoustic model and the neural vocoder. Section 4 presents the results of the subjective evaluation conducted by the challenge organizers. A conclusion is give in the end..

2. System Description

2.1. Data

The Blizzard Challenge 2021 data consists of five hours of European Spanish (ES) speech from one female native speaker. Sentences are clearly spoken and the speech files consist of noise-free studio quality recordings in WAV format sampled at 48 kHz, 16 bits, mono.

In order to enrich the training material, external data comprising Brazilian Portuguese (BP) and American English (AE) was included in our experiments.

For BP, a proprietary multi-speaker dataset, with 30 hours of speech from 2 male and 4 female speakers was used. Recordings consist of phonetically rich sentences sampled at 22kHz, whose file format and recording conditions are similar to those of the ES dataset.

With regard to AE, a subset of LibriTTS [11] was selected with characteristics similar to those of the BP data in terms of gender balance and amount of data, with 30 hours of speech from 22 male and 42 female speakers. LibriTTS is originally sampled at 24 kHz.

The sample rate used in our experiments was 22kHz, so speech files were downsampled whenever necessary.

As will be detailed later, all three languages were used in our experiments, but the final version submitted for evaluation relied solely on the ES and BP data.

2.2. Linguistic Front-end

The linguistic front-end transforms the input text sequence into a phoneme sequence to be fed to the acoustic model. It consists of a text normalizer, which expands numbers, acronyms and non-linguistic symbols, followed by a grapheme-to-phoneme (G2P) converter. We employ three independent front-ends, one for each language present in our data, i.e. European Spanish (ES), Brazilian Portuguese (BP) and American English (AE). All three G2P modules work with a common proprietary phoneme set comprising (BP: 42 phonemes + 7 pause types; ES: 26 phonemes + 7 pause types; AE: 29 phonemes + 7 pause types). It is important to notice that, in such phoneme set, 84.6 percent of the ES phonemes are also present in BP, while only 61.5 percent of the ES phonemes are present in AE.

The implementation of G2P for ES and BP is based on a set of language specific rules, since, in most cases, pronunciation of words can be inferred from their spelling. Additionally, a lookup table is consulted in the case of words for which the rules fail. In the case of AE, the primary way to obtain the correct pronunciation is through a pronunciation dictionary, while out of vocabulary words are handled by a decision tree. In all cases, post-processing rules are applied to deal with co-articulation.

2.3. Acoustic Model

For the acoustic model we adopt the Tacotron2 [12] attention-based encoder-decoder architecture. The encoder consists of three CNN layers followed by one bi-directional LSTM layer. We use phoneme sequence as input to prevent the model from learning pronunciation patterns straight from the character sequence. A language label is fed to the encoder via a non-linguistic token preceding the phoneme sequence, so the network can be conditioned on the target language.

The network is also conditioned on the speaker identity by a 512 dimensional x-vector [13] added to the encoder output [14]. X-vectors are computed during training by an independent pre-trained extractor (in this work we use the Kaldi SITW model[1]. Since it is a 16 kHz extractor, audio files are downsampled to 16kHz before computing x-vectors). At inference time, the mean x-vector of the target speaker is used to condition the network.

The network uses location sensitive attention with guided attention loss to perform a soft alignment between the phoneme sequence and the output mel-spectrogram, by determining the most related context at each step with respect to the previous decoding output and the encoder output states.

The decoder is comprised of a two layer LSTM network, a fully connected PreNet and a convolutional PostNet. The PreNet consists of two fully connected linear layers, while the PostNet is constructed by five convolutional layers to reduce the information loss caused by the unidirectional decoding characteristic of the decoder. A stop token is used to predict the end of the decoding process.

The output of the acoustic model is a sequence of 80-dim mel-spectrogram values.

2.4. Vocoder

We use Parallel Wavegan [15] to process the output of the acoustic model. It is a parallel waveform generation method based on a generative adversarial network (GAN), that can generate high quality speech samples from the mel-spectrogram values, with small footprint and a reasonable computation time. The Parallel Wavegan architecture consists of a non-causal WaveNet generator and a single CNN discriminator (D). Based on GANs, the generator learns a distribution of realistic waveforms by trying to deceive the discriminator into recognizing the generated samples as real. Moreover, the discriminator is trained to correctly classify the generated sample as fake while classifying the ground truth as real. By combining adversarial training with an auxiliary multi-resolution short-time Fourier transform (STFT) loss function, Parallel WaveGAN learns the time-frequency characteristics of realistic speech efficiently.

3. Experiments / System building

3.1. Data preparation

No relevant pronunciation errors could be found in the material provided by the challenge organizers. The SNR ratio of the speech samples was high, and no significant distortion such as clipping, reverberation, etc, was observed. Yet, some pre-processing steps were necessary to ensure proper training of the system modules, as described below.

Long sentences were broken to ensure the duration of training samples would be a maximum of 10 secs. Whenever possible, punctuation symbols were used as break points.

[1] https://kaldi-asr.org/models/m8

Sentences were classified according to the last punctuation character. This was necessary to create separate training sets for each sentence type, as will be detailed in section 3.2.

In the case of sentences ending with a question mark (?) three categories were considered:

- sentences in which the first word is an interrogative pronoun (e.g "quién", "cuándo" "cómo") , or in which the second or third word is an interrogative pronoun preceded by a preposition, were classified as **"open-questions"**.
- sentences in which the word *"o"* occurs were classified as **"or-questions"**
- the remaining interrogative sentences were classified as **"yes-no-questions"**

If a question could be classified in more than one category, then it was marked as an **"open-question"**.

Additionally, two non-interrogative classes were considered:

- sentences ending with an exclamation mark (!) were classified as **"exclamatory"**.
- all remaining sentences were classified as **"declarative"**

Informal tests showed that the application of the aforementioned rules resulted in very few classification errors.

Speech files were converted to WAV format sampled at 22kHz, 16 bits, mono. Leading and trailing silences of all speech samples were normalized between 200 and 300 ms. Additionally, a 10 Hz low pass filter was applied to all samples in order to remove DC level.

3.2. Acoustic model training

The acoustic model was built with the espnet [16] toolkit[2]. More specifically, the refactored version, called espnet2 [17], was used. In our experiments we employed a dynamic batch size, with number of bins equal to 2560000.

An initial acoustic model trained with only the 5h Spanish data was created as a baseline. The hypothesis we wished to validate was whether a better model could be obtained without requiring any additional data in the target language. For that, we boosted the original model by augmenting the training data with 30 hours of a second language. We used the multispeaker BP and the AE data described in section 2.1 to train two separate ES+BP and ES+AE models. These were trained for 500 epochs, with 200 iterations per epoch.

Regarding phonetics, ES is much closer to BP than to AE, so we wanted to assess to what extent the choice of the supplemental language would affect the quality of the boosted models.

We performed AB-like listening tests with 6 expert listeners in order to compare the abovementioned models. A text corpus consisting of phonetically rich sentences, proper names, addresses and currency values were used in the listening tests[3]. Besides indicating their preference, listeners were asked to give an overall impression about the degradations they could perceive in the synthetic speech, focusing on naturalness and intelligibility. Similarity was not evaluated.

Table 1 shows the result of the AB tests comparing the three models. 100% of the listeners chose the ES+BP over both the ES-only and ES+AE models, while 50% of the listeners chose the ES+AE model over the ES-only model. The most relevant overall impressions reported by the listeners were:

Table 1: *The AB test result for different acoustic models trained with different datasets: European Spanish (ES); European Spanish + Brazilian Portuguese (ES+BP); European Spanish + American English (ES+AE).*

Model Eval	ES+BP	ES+AE	ES	No preference
ES+BP vs ES	100%	-	0%	0%
ES+BP vs ES+AE	100%	0%	-	0%
ES+AE vs ES	-	50%	16.7%	33.3%

- ES-only model: Some spectral noise and unnatural artifacts were reported. Despite a few pronunciation errors, no gross errors such as muffling, skipping, repetitions and early stops were perceived. This is a surprisingly good result given that only 5 hours of training data were used. Our previous experience with BP showed that at least 15 hours of training data was necessary to train an acoustic model with acceptable quality. We hypothesize that ES is more regular than BP with regard to the grapheme to phoneme mapping, in addition to having a less diverse set of phonemes, which might help achieving a good quality with less training data.
- ES+AE model: augmenting the training data with AE resulted in a less quavering voice and less spectral noise. Unnatural artifacts were reduced but not eliminated. On the other hand, the quality of the prosody was considered worse, since the synthetic sentences had a flatter and sometimes unnatural intonation.
- ES+BP model: training enriched with BP data improved the naturalness of prosody, eliminated most of the unnatural artifacts and improved the general audio quality (less quavering voice / less spectral noise).

Therefore, AE data was no longer used in subsequent experiments.

The ES+BP model was later fine-tuned with only the ES data, in order to better fit it to the target speaker. 500 epochs were used with 200 iterations per epoch for this task.

The resulting model achieved high naturalness, intelligibility and similarity with the target speaker's voice. However, it was noted that non-declarative sentences were not properly modeled. In such cases, the final punctuation was usually ignored and the sentence had a declarative intonation pattern. In order to overcome this limitation, specialized models for each of the non-declarative sentence types described in Section 3.1 were trained. Such models were obtained with a second round of fine-tuning, in which the training data consisted exclusively of sentences of the target type. In this second round of fine-tuning 300 epochs were executed, with 20 iterations per epoch.

205 sentences extracted from the ES dataset were used to fine-tune the **"open-questions"** model, and 214 sentences, also extracted from the ES data, were used to adapt the **"yes-no-questions"** model. In the case of the **"or-questions"** only 14 sentences existed in the ES dataset, which proved to be insufficient to adapt the model. For this reason, 185 *or-questions* were selected from the BP dataset and added to the adaptation data. Similarly, the adaptation data for the **"exclamatory"** model consisted of 71 sentences from the ES dataset and 129 from the BP dataset.

In the case of the declarative sentences no second round of fine-tuning was executed.

[2] https://github.com/espnet/espnet

[3] Audio samples are available at https://bit.ly/38QL9BY

3.3. Vocoder training

The neural vocoder was built using Hayashi's publicly available unofficial implementation[4] of Parallel Wavegan on GitHub. To avoid training the neural vocoder from scratch, a model pre-trained with 3000k steps on the LJSpeech [18] dataset was used to initialize the network weights. Such model is publicly available at the Parallel Wavegan repository on GitHub.

600k training steps were executed wit the same data used by the base acoustic model (i.e. 5 hours ES + 30 hours BP) in order to obtain a multi-speaker neural vocoder. This multi-speaker model was later fine-tuned with only the target speaker data, so the model could specialize in the voice of the target speaker. In this case, 1500k training steps were executed. A better perceptual result was achieved when the weights obtained in steps 1300k, 1400k and 1500k were averaged in order to get the final model weights.

3.4. Inference

At inference time, each sentence was synthesized independently. A sentence classifier, based on the rules described in Section 3.1, was used in order to select the most appropriate acoustic model for each sentence.

The acoustic model was conditioned on the phonetic sequence generated by the linguistic front-end, as well as on the speaker ID and the target language. The mean x-vector of the target speaker was used to inform the speaker ID, while a text label preceding the phoneme sequence was used to inform the target language.

4. Results

This section discusses the evaluation results of our system on the Blizzard Challenge 2021's hub task. Our system is identified as **G**, whereas system **R** is natural speech and systems A/B/C/D/E/F/I/J/K/L/N are the the remaining 11 participating teams.

The evaluation was conducted online. The order of presentation of the systems was varied according to a Latin Square design. The assessment consisted of independent test sessions, whose ultimate goal was to generate an assessment metric for each of the following aspects:

- naturalness (MOS): a score representing how natural or unnatural the sentence sounded on a scale ranging from 1 [Completely Unnatural] to 5 [Completely Natural].
- similarity to original speaker: a score representing the degree of similarity between the synthesized voice and the target speaker's voice on a scale ranging from 1 [Sounds like a totally different person] to 5 [Sounds like exactly the same person].
- intelligibility: a word error rate (WER) count, obtained by asking participants to listen to synthetic utterances and then type in what they just heard.

Three types of listeners took part in the evaluation: paid listeners (native Spanish speakers), speech experts and volunteers (non-experts).

4.1. Naturalness

Figure 1 shows the boxplot for naturalness considering ratings from all listeners. Among all systems submitted to the challenge, only one managed to obtain a better result than ours, while three other systems obtained comparable results.

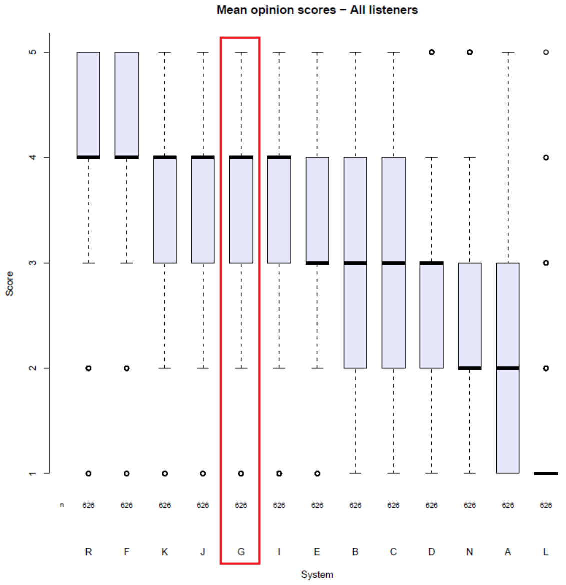

Figure 1: *Mean Opinion Scores (MOS) for all listeners - Overall Impression.*

4.2. Similarity

The speaker similarity boxplot with mean opinion scores from all listeners is shown in Figure 2. As with the naturalness test, our system outperformed most of the other systems. Two participants obtained better results, while a third participant achieved a rating that can be considered statistically similar to ours.

4.3. Intelligibility

The intelligibility analysis was separated into two sections. One featured natural sentences from the Sharvard corpus. Another called SUS contained sentences provided by TALP-UPC[5] and Aholab-EHU[6] research laboratories, manually generated containing different syntactic structures and lexicon requirements for semantically unpredictable sentences.

Figure 3 shows the result for the word error rate on Sharvard sentences, while Figure 4 shows the result for the word error rate on SUS sentences. In both cases the results obtained by our system can't be considered statistically different from those obtained by most participants, including the top ranked.

4.4. Discussion

One point that must be highlighted is that the authors of this paper have access to a large amount of Latin American Spanish speech data that could have been used to obtain better results for the Blizzard Challenge. Despite that, we decided not to use such data in our submission, since our main objective was to evaluate if a high quality system could be obtained without making use of any additional data in the target language. We also wanted to asses whether adding data in a different but similar language could improve the results.

[4] https://github.com/kan-bayashi/ParallelWaveGAN

[5] http://www.talp.upc.edu/

[6] https://aholab.ehu.eus/

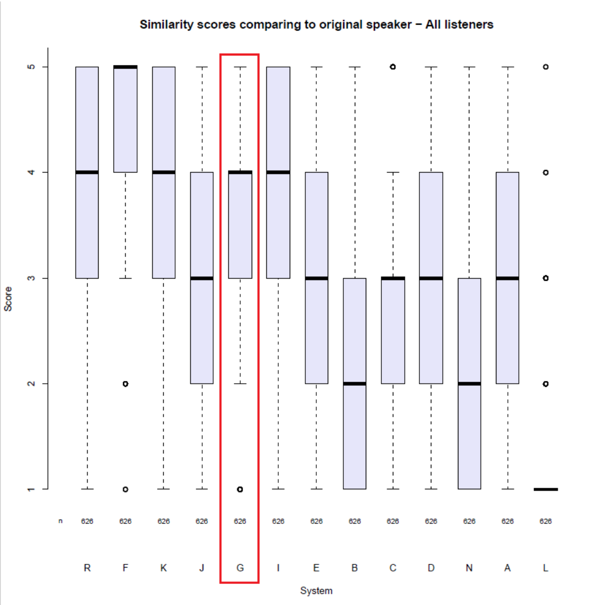

Figure 2: *Similarity scores comparing to original speaker for all listeners - Overall Impression.*

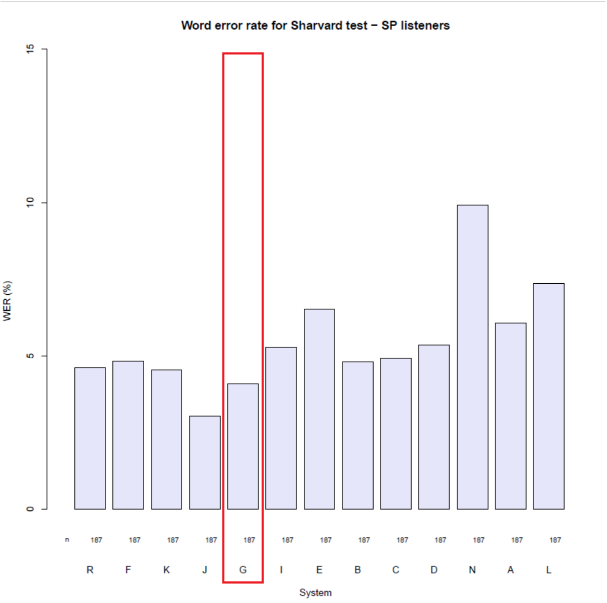

Figure 3: *Word error rate for Sharvard test for native speakers of Spanish.*

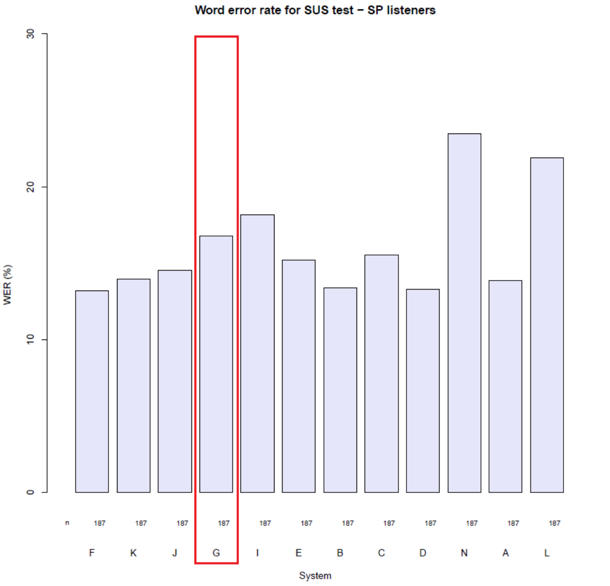

Figure 4: *Word error rate for SUS test for native speakers of Spanish.*

During system development, other strategies to improve the prosody of non-declarative sentences were tested, such as (1) insertion of a non-linguistic label in the beginning of each sentence to inform the sentence type; (2) concatenation of a one-hot embedding to the encoder output, and (3) utilization of Global Style Tokens [19] to capture the different intonation patterns in an unsupervised way. All of these approaches resulted in worse results than those obtained by our submitted system, especially in the case of sentence types that were underrepresented in the training data. Despite that, we believe such strategies deserve further investigation.

5. Conclusions

In this paper we have presented the joint submission from CPQD and Unicamp (CPQD-Unicamp) to Blizzard Speech Synthesis challenge 2021. The system consists of a bilingual linguistic front-end, an acoustic model based on Tacotron2 and a Parallel Wavegan neural vocoder. A multi-speaker Brazilian Portuguese dataset was added to the Blizzard 2021 dataset in order to train a bilingual acoustic model, while the vocoder was trained with data in European Spanish, Brazilian Portuguese and American English. Such modules were later fine-tuned with the target speaker data, and specialized models have been trained to better handle intonation patterns related to the punctuation mark. The evaluation results indicate that enriching the acoustic model training with data from a second language can improve the model quality.

6. References

[1] "The blizzard challenge website." [Online]. Available: https://www.synsig.org/index.php/Blizzard_Challenge

[2] "The blizzard challenge 2021 website." [Online]. Available: https://www.synsig.org/index.php/Blizzard_Challenge_2021

[3] Z. Wu, Z. Xie, and S. King, "The blizzard challenge 2019," in *Proc. Blizzard Challenge Workshop*, vol. 2019, 2019.

[4] X. Zhou, Z.-H. Ling, and S. King, "The blizzard challenge 2020,"

in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 1–18.

[5] J. Latorre, J. Lachowicz, J. Lorenzo-Trueba, T. Merritt, T. Drugman, S. Ronanki, and V. Klimkov, "Effect of data reduction on sequence-to-sequence neural tts," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7075–7079.

[6] H.-T. Luong, X. Wang, J. Yamagishi, and N. Nishizawa, "Training Multi-Speaker Neural Text-to-Speech Systems Using Speaker-Imbalanced Speech Corpora," in *Proc. Interspeech 2019*, 2019, pp. 1303–1307.

[7] Y. Deng, L. He, and F. Soong, "Modeling multi-speaker latent space to improve neural tts: Quick enrolling new speaker and enhancing premium voice," 2019.

[8] Q. Yu, P. Liu, Z. Wu, S. K. Ang, H. Meng, and L. Cai, "Learning cross-lingual information with multilingual blstm for speech synthesis of low-resource languages," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5545–5549.

[9] K. R. Prajwal and C. V. Jawahar, "Data-efficient training strategies for neural tts systems," in *8th ACM IKDD CODS and 26th COMAD*, ser. CODS COMAD 2021. New York, NY, USA: Association for Computing Machinery, 2021, p. 223–227. [Online]. Available: https://doi.org/10.1145/3430984.3431034

[10] J. Latorre, C. Bailleul, T. Morrill, A. Conkie, and Y. Stylianou, "Combining speakers of multiple languages to improve quality of neural voices," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 37–42.

[11] H. Zen, R. Clark, R. J. Weiss, V. Dang, Y. Jia, Y. Wu, Y. Zhang, and Z. Chen, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech*, 2019. [Online]. Available: https://arxiv.org/abs/1904.02882

[12] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[14] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *arXiv preprint arXiv:1806.04558*, 2018.

[15] 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[16] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "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)*. IEEE, 2020, pp. 7654–7658.

[17] S. Watanabe, F. Boyer, X. Chang, P. Guo, T. Hayashi, Y. Higuchi, T. Hori, W.-C. Huang, H. Inaguma, N. Kamo, K. Shigeki, C. Li, J. Shi, A. S. Subramanian, and W. Zhang, "The 2020 espnet update: New features, broadened applications, performance improvements, and future plans," in *IEEE Data Science and Learning Workshop (DSLW)*, Jun. 2021. [Online]. Available: https://www.merl.com/publications/TR2021-073

[18] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[19] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.