The SCUT Text-To-Speech System for the Blizzard Challenge 2021

Weiheng Liu[1]*, Yitao Yang*[1]*, Jiangwei Li*[1]*, Jinghui Zhong*[1]

[1]School of Computer Science & Engineering, South China University of Technology

ryanliu_scut@163.com, ytyang_scut@163.com, csjwli@mail.scut.edu.cn, jinghuizhong@scut.edu.cn

Abstract

In this paper, we present our solution for the Blizzard Challenge 2021 Spoke task, which is to build a code-switched speech synthesis system for European Spanish and English with only Spanish dataset. The major challenges of code-switched text are language-independent representation of linguistic information and cross-language speaker transfer. For these difficulties, a set of phonological embedding derived from the International Phonetic Alphabet(IPA) is applied to uniformly identify bilingual texts and facilitate knowledge sharing among multiple languages. Meanwhile, our system uses predefined speaker embedding to control the voice of the generated speech. In addition, we introduced a variational autoencoder to extract hidden features in speech in order to balance the data differences between multiple datasets. The results of the evaluation have demonstrated the effectiveness of our method in code-switched speech synthesis.

Index Terms: Blizzard Challenge 2021, speech synthesis, Code-switched speech synthesis, Bilingual speech synthesis

1. Introduction

The Blizzard Challenge is an annual scientific event devised to understand the research status in the speech synthesis communities by comparing various speech synthesis techniques on a common dataset. The challenges this year are arranged for European Spanish speech synthesis, including (1) Hub task to build a Spanish voice given speech data. (2) Spoke task to build a Spanish and English bilingual speech synthesis system with only a Spanish speech dataset. We select the Spoke task to construct a bilingual synthesis system, which involves code-switched speech synthesis on only monolingual corpus.

Recently, the speech synthesis systems with neural network-based acoustic models and vocoders have achieved great success[1, 2, 3, 4, 5, 6]. Their generated speech can even be comparable to human voices in naturalness and fidelity. However, code-switching utterances, consisting of words in multiple languages, are still a tough challenge for the speech synthesis system. As early as the last decade, HMM-based speech synthesis systems[7, 8] have straightforwardly used bilingual databases to build code-switched speech synthesis systems. But bilingual databases are scarce and manual collection is expensive. At present, most of the abundant speech resources are monolingual corpus. Therefore, how to transfer the knowledge learned from monolingual corpus to other languages has became a hot topic in the research field. The voice transfer between languages is intuitive because humans share the same vocal organs, and the vocalization of fine particle units has similarities in various languages. The mainstream solution is to extract the language-independent speaker embedding to disentangle the speaker information and language information[9, 10, 11, 12, 13, 14, 15, 16, 17, 18]. In [9], adversarial loss is used to decouple text information and speaker information. Also, the explicit labeling of languages and speakers also facilitates information disentanglement[17, 18].

At the same time, the different phoneme sets between different languages increase the difficulty of code-switched synthesis. Therefore, the International Phonetic Alphabet (IPA) [14]or the phonological features derived from the International Phonetic Alphabet[13] are used as uniform representations of different language texts, leading to knowledge sharing between languages. Another explicit representation for various languages is Byte[15], while some methods consider the unified representation of phonemes as learnable parameters[12, 11]. Meanwhile, Phonetic PosteriorGrams (PPGs) extracted from a speaker-independent speech recognition model are deemed speaker-independent and language-independent and can be unified features to support arbitrary texts in multiple languages[10].

There is also a more straightforward method to solve the data shortage issue. [19, 20] conducts cross-language voice conversion to generate the voice of a specific speaker in an unpaired language and this kind of data augmentation method has also achieved good performance.

Due to the limited data of the Spoke task, based on [13], we manually designed the representation of the phonological features in Spanish and English to complete the code-switched synthesis. For cross-lingual speaker transfer, awing to the data restriction, we only added a one-hot speaker embedding as the input of the acoustic model. Finally, to avoid the training instability caused by implicit differences between multiple monolingual datasets, we also included an additional information extractor implemented as variational autoencoder (VAE), which has been verified to extract hidden information such as prosody in speech[21][22][23]. For the vocoder part, we used Multi-band WaveRNN[24], a variant of the WaveRNN vocoder[2], which predicts samples for multiple subbands simultaneously and all samples from subbands are summed up to restore the original waveform by the synthesis filter. This subband-generation method enables Multi-band WaveRNN to contain both high generation speed and high fidelity.

The rest of this paper is organized as follows: First, we describe our method in section 2 and the experiment details will be presented in section 3. Then, section 4 shows the evaluation results and there are some discussions about our implementation's weak points. Finally, we will conclude our paper in section 5.

2. Method

2.1. Overall Architecture

Since we choose the Spoke Tasks(SS1) in which the synthesis texts will contain some English words, we hope to build a system that supports bilingual speech synthesis. In the front-end module, the phonological features extraction method is

Table 1: *Dataset information*

dataset	speakers	utterances	duration/hrs
BC2021	1	4883	5.24
Common Voice	401	56243	63.2
CSS10 Spanish	3	8188	14.7
VCTK	61	21683	6.0
Total	466	90997	96.3

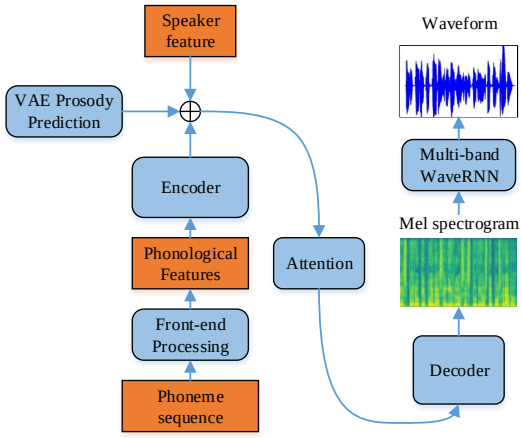

Figure 1: *Overall Architecture*

adopted, which unifies the representation of Spanish and English phonemes. Besides, we use VAE to capture the prosody features and one-hot vectors for speaker identification.

The model architecture is shown in Figure 1. We choose the Tacotron2[5] as the acoustic model, which has an encoder-decoder architecture. Specifically, the encoder extracts the linguistic features from the phoneme sequence, while the decoder outputs the frame-level Mel-spectrogram which is aligned to the corresponding phoneme context by attention mechanism. Finally, Multi-band WaveRNN [24], the vocoder, restores the Mel-spectrograms to the waveform.

2.2. Data Selection

The Blizzard Challenge committee provides an approximately 5-hour Spanish dataset of a female speaker containing texts and speech data. These materials are insufficient to modeling the Spanish linguistic features, so we externally utilize the following materials:

- Common Voice[25]: where we pick up the Spanish belong to European area, including Sur peninsular, Centro-sur peninsular and Norte peninsular.
- CSS10[26]: only choose the Spanish subset.

In terms of modeling the English one, the VCTK[27] dataset is used. We only select the utterances that belong to female speakers. Some silence or low volume wave files are deleted. All wave files are resampled to 22.05kHz and trimmed the heading and tailing silence. Detailed information refers to Table 1.

2.3. Front-end processing

This module aims to convert Spanish and English characters to uniform phoneme representations. Firstly, the texts are translated into International Pronounce Alphabet (IPA) by using the espeak-ng tool. It is worth mentioning that Spanish and English words in bilingual texts need to be identified and translated separately. In the training stage, because we use several monolingual corpus but not bilingual corpus as training materials, the identification of Spanish and English words in single sentences can be omitted. But for the bilingual testing sentences, we apply the Spanish dictionary and English dictionary of LibreOffice to distinguish between Spanish words and English words in a sentence. We believe that native Spanish speakers tend to use Spanish to pronounce words that exist in both Spanish and English, so the principle we use is that Spanish words take precedence. It means that one word will be marked as an English word when it only appears in the English dictionary and not in the Spanish dictionary.

Secondly, to support the code-switching synthesis, phonological features are introduced. We design an encoding method of phonological features, which are listed in Table 2. 128 phonemes, including 126 ones from IPA and 2 added silence symbols, are encoded into 55-dimensional one-hot vectors according to their phonological features, such as vowel/consonant, vowel frontness, consonant place, etc. For example, the feature "vowel frontness" consists of five classes: "central", "front central", "central back", "back", "front", and "default". So we use 5 dimensions in the 55-dim binary vector to describe this feature. Since the "vowel frontness" of phoneme "*a*" is "front", the dimension corresponding to "front" is set to "1". Thus, out-of-sample(OOS) phonemes could be inferred from the existed IPA based on their similarity in the embedding space.

Practically, the phoneme *'sil'* is important to modeling the pause between words and sentences, which nearly appears in every utterance. Not only will it significantly affect the naturalness of synthesis speech, but also help the alignment of the attention mechanism.

2.4. Speaker Identification

To achieve the ability of voice conversion, the speaker identification feature is indispensable. Because of the limited external speech data, the common speaker identification models pretrained by a huge amount of data are unavailable, such as i-vector[28] and x-vector[29]. We represent them with one-hot vectors as a rough substitute. The one-hot vectors are projected to 512-dimensional embeddings, which will automatically learn the speaker feature during the training phase.

2.5. VAE prosody extraction

Prosody extraction plays an important role in high naturalness synthesis. Zhang. et al.[21] firstly use the VAE to predict the Global Style Token(GST)[30]. Generally, Mel-spectrogram features will be firstly input to a reference encoder including six 2-D convolutional layers and two GRU layers. Then two fully connected layers follow to it, which generate the mean μ and standard deviation σ of latent variables $\mathbf{z}$ respectively. The $\mathbf{z}$ will be finally sampled with reparameterization trick and subsequently added to the encoder output. The Tacotron, which acts as a decoder of autoencoder, reconstructs the Mel-spectrograms from the combined encoder states. The loss function will compute a variational lower bound which consists of a reconstruc-

Table 2: *Phonological features and their classes*

feature	classes
vowel or consonant	vowel, consonant
VUV	voiced, unvoiced
vowel frontness	central, front central, central back, back, front, default
vowel openness	mid, cross-mid, close, open-mid open, open, close close-mid, open-mid, default
vowel roundedness	unrounded, rounded, default
stress	primary stress, secondary stress, unstressed, default
consonant place	bilabial, alveolar, labiodental, retroflex, postalveolar, uvular, glottal, velar, palatal, dental, labial-velar
consonant manner	nasal, affricate, lateral-approximant, tap, approximant, stop, trill, default
diacritic	nasalized, rhoticity, syllabic, velarized, default
length	long, default
sil	sil
EOS	EOS

tion loss term and a KL loss term, as equation 1. In inference, we could sample a $\mathbf{z}$ from an isotropic multivariate Gaussian distribution $\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$ and the model would predict the Mel-spectrograms with special style if appropriately trained.

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - KL[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})] \tag{1}$$

2.6. End-to-end speech synthesis model

2.6.1. Encoder

After the front-end processing, phonological features are first transformed into 512-dimensional embeddings by a linear layer. Then they are passed through a stack of three 1-D convolution layers, each of which has 512 filters with kernel size 5. Subsequently, it is followed by a single Bi-LSTM layer, whose hidden size is 256 in each direction. Also, the prosody features and the speaker features are projected to 512-dimensional hidden representations respectively. Finally, those three features will be summed up together as the combined encoder outputs. The output sequence would be the same length as the input phoneme sequence.

2.6.2. Decoder

The decoder consists of a pre-processing net, an attention based RNN net and a post-processsing net. The pre-processing net contains 2 fully connected layers. It takes the prediction of previous time step, acting as a bottleneck to enhance the model's generalization. The RNN net consists of a stack of 2 LSTM layers with 1024 units, which takes the concatenation of the aligned context from the attention module and the output of the pre-processing net. The output of LSTM is projected through two linear transforms respectively. One is to predict the target Mel-spectrogram, the other is to predict the probability of completion of the output sequence. Finally, The 5-layer convolutional post-processing layer improves the effect of overall Mel-spectrogram reconstruction.

To speed up the training, we set the reduction factor of Tacotron to 3, i.e., each decoder step generates 3 frames of Mel-spectrogram. Even though it would sacrifice some quality, it exchanges for a bigger batch size setting and easier alignment.

2.7. Multi-band WaveRNN Vocoder

The vocoder inverts the Mel-spectrogram feature into a time-domain waveform. For the purpose of reducing the computational cost and speeding up the inference of vocoder, multi-band parallel strategy is introduced[24]. Generally speaking, every frame of Mel-spectrogram is decomposed to several sub-bands by Pseudo quadrature mirror filter banks(PQMF). And every subband is downsampled by the factor of N (the number of frequency bands), thus the computation cost will be reduced. The modified WaveRNN could accept the previous predicted sample with all subbands as input and predict the next sample in multiple subbands simultaneously. After inference, the signals from multiple bands will be restored to waveform by the synthesis filters.

3. Experiment

3.1. Training

We train the aforementioned end-to-end model utilizing all the data mentioned above. Because of the small amount of the BC2021 dataset and the demand for bilingual synthesis, external English and Spanish corpus are needed for model to adequately learn the linguistic features. In terms of the training of the VAE, we utilize the annealing trick mentioned in[21, 31] to avoid the KL collapse problem. The training of Multi-band WaveRNN follows the principle of pretrain-and-finetune. During the pretrain phase, WaveRNN is trained being fed the whole dataset. Since the generated Mel-spectrogram has some loss, we intend to enable WaveRNN to fix the gap between ground truth Mel-spectrograms and synthesis ones. We finetune it by inputing the predicted Mel-spectrograms from Tacotron among the BC2021 dataset.

3.2. Inference

The BC2021 committee releases 224 English-Spanish-switched texts to synthesis for evaluation. We set the speaker feature to whom from the BC2021 dataset. Besides, we use the VAE module to predict a mean and a standard deviation of each utterance and compute the averaged mean and averaged standard variance of the whole BC2021 dataset, which will be set as the prosody feature during the inference. Finally, texts are transformed into Mel-spectrograms with Tacotron and the WaveRNN generates time-domain waveforms. The final results are thus obtained.

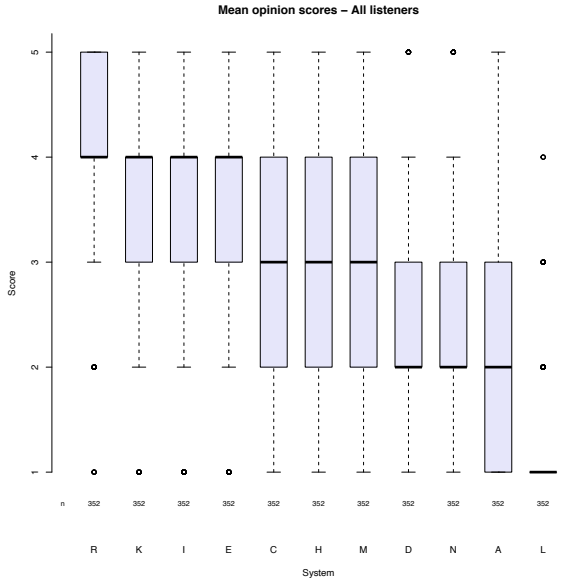

Figure 2: *Mean Opinion Scores (naturalness)*

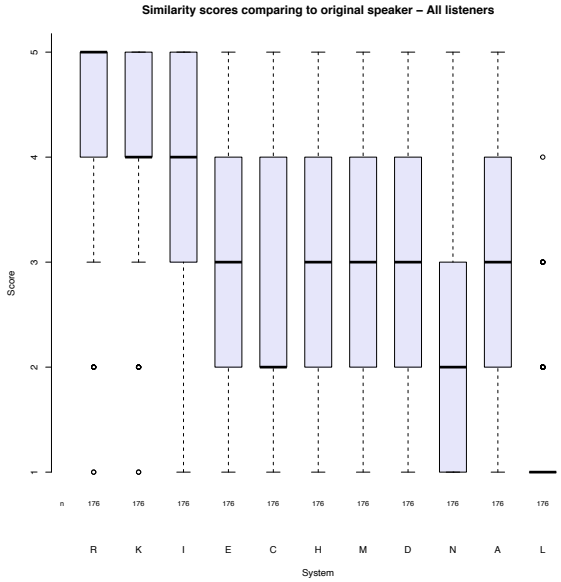

Figure 3: *Mean Opinion Scores (similarity to original speaker)*

4. Results

4.1. Evaluation

The subject evaluation of Spoke task assesses our performance from three aspects: naturalness, similarity to original speaker and acceptability of English words. And the mean opinion score (MOS) ranges from 1 (bad) to 5 (excellent) expresses the grades of each aspect. The naturalness represents how natural or unnatural the sentence sounded and the similarity to original speaker is to check whether the voice of the sentence deviates from that of target speaker. Meanwhile, in order to evaluate the effect of code-switched synthesis, the acceptance of English words was contained to evaluate the pronunciation of English words mixed in Spanish sentences.

The MOS results of naturalness, similarity to original speaker and acceptability of English words are shown in Figure 2, Figure 3, and Figure 4, respectively. The symbol M represents our system. According to the results, our performance is in the middle of all participating teams. In terms of naturalness, our MOS score is 2.72, which means our generations are between natural and unnatural, indicating that our system still needs improvement. Also, the speaker similarity of our system is not eye-catching enough. Finally, in terms of code-switched generation, our system got an MOS score of 2.30. It reveals that our system can achieve limited code-switch synthesis and our unified phonological representation helps knowledge sharing between different languages.

4.2. Discussion

Next we carry out some analysis of the system. From the evaluation results, it can be seen that our synthesis system can generate understandable speech, and the phonological front-end does make the system capable of code-switched synthesis. However, the performance is lower than expected. We attribute the fault mainly to VAE's posterior collapse. During our training process, the output of our reference encoder collapsed into a standard normal distribution, which caused the reference encoder's failure to balance the implicit features of every sentence in all datasets. Under this situation, the onehot speaker embedding we input acts as an identification of not only the speaker but also the database. So the Tacotron model is just like an average model among the corresponding training dataset when the speaker is fixed. Therefore, it is easy to cause unnatural pauses, repeats, missed pronunciation, etc., due to the unstable attention mechanism of the average model. Moreover, since our speaker embedding is not lingual-independent, it also involves the style features of the corresponding dataset. So, when generating the final code-switched speech, the pronunciation corresponding to the English phoneme will include Spanish style, resulting in the English voice with Spanish accent.

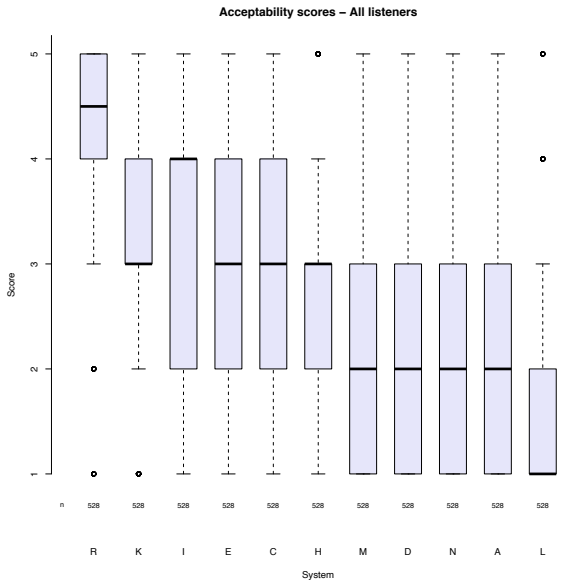

Figure 4: *Mean Opinion Scores (acceptability of English words)*

5. Conclusions

In this article, we describe our SCUT system in detail. Our overall architecture is based on Tacotron 2 and Multiband WaveRNN. For the Spoke task, which is about code-switched synthesis, we absorbed the zero-shot method in [2], and manually formulated the phonological embedding from the Spanish and English IPA, so that the Spanish and English texts were uniformly represented. In this way, the model can learn the pronunciation of Spanish and English at the same time, even in scenarios with only monolingual datasets. In addition, we use speaker embedding to distinguish the voice in order to complete cross-language speaker transfer. Also, we try to utilize VAE to balance the implicit differences between multiple monolingual datasets, but we fell into the KL-vanishing problem, leading to our system's unsatisfactory performance. In future work, we will pay more attention to speaker transfer and bilingual synthesis in low-resource scenarios.

6. Acknowledgements

This work is supported by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No.: 2017ZT07X183)

7. References

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," pp. 1–15, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[2] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimber, A. Van Den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *35th International Conference on Machine Learning, ICML 2018*, vol. 6, pp. 3775–3784, 2018.

[3] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based Recurrent Neural Networks," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, no. September, pp. 1964–1968, 2014.

[4] S. O. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, no. Nips, pp. 2963–2971, 2017.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, pp. 4779–4783, 2018.

[6] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, and Q. Le, "Tacotron: Towards end-To-end speech synthesis," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, pp. 4006–4010, 2017.

[7] S. Sitaram, S. K. Rallabandi, S. Rijhwani, and A. W. Black, "Experiments with Cross-lingual Systems for Synthesis of Code-Mixed Text," pp. 76–81, 2016.

[8] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner, "From multilingual to polyglot speech synthesis," *Proc Eurospeech*, 1999.

[9] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. J. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September, pp. 2080–2084, 2019.

[10] Y. Cao, S. Liu, X. Wu, S. Kang, P. Liu, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Code-switched speech synthesis using bilingual phonetic posteriorgram with only monolingual corpora," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7619–7623.

[11] Y. J. Chen, T. Tu, C. C. Yeh, and H. Y. Lee, "End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September, no. 2, pp. 2075–2079, 2019.

[12] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T. Y. Liu, "LRSpeech: Extremely Low-Resource Speech Synthesis and Recognition," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2802–2812, 2020.

[13] M. Staib, T. H. Teh, A. Torresquintero, D. S. Ram Mohan, L. Foglianti, R. Lenain, and J. Gao, "Phonological features for 0-shot multilingual speech synthesis," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-October, pp. 2942–2946, 2020.

[14] H. Hemati and D. Borth, "Using IPA-Based Tacotron for Data Efficient Cross-Lingual Speaker Adaptation and Pronunciation Enhancement," 2020. [Online]. Available: http://arxiv.org/abs/2011.06392

[15] M. He, J. Yang, and L. He, "Multilingual Byte2Speech Text-To-Speech Models Are Few-shot Spoken Language Learners," 2021. [Online]. Available: http://arxiv.org/abs/2103.03541

[16] L. Xue, W. Song, G. Xu, L. Xie, and Z. Wu, "Building a mixed-lingual neural TTS system with only monolingual data," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September, pp. 2060–2064, 2019.

[17] J. Yang and L. He, "Towards universal text-to-speech," in *INTERSPEECH*, 2020.

[18] Z. Cai, Y. Yang, and M. Li, "Cross-lingual Multispeaker Text-to-Speech under Limited-Data Scenario," 2020. [Online]. Available: http://arxiv.org/abs/2005.10441

[19] S. Zhao, T. H. Nguyen, H. Wang, and B. Ma, "Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-October, pp. 2927–2931, 2020.

[20] G. Huybrechts, T. Merritt, G. Comini, B. Perz, R. Shah, and J. Lorenzo-Trueba, "Low-resource expressive text-to-speech using data augmentation," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6593–6597, 2021.

[21] Y. J. Zhang, S. Pan, L. He, and Z. H. Ling, "Learning Latent Representations for Style Control and Transfer in End-to-end Speech Synthesis," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2019-May, pp. 6945–6949, 2019.

[22] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-Hierarchical Fine-Grained Prosody Modeling for Interpretable Speech Synthesis," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2020-May, pp. 6264–6268, 2020.

[23] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, pp. 4440–4444, 2019.

[24] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, D. Su, and D. Yu, "DurIAN: Duration informed attention network for speech synthesis," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-October, pp. 2027–2031, 2020.

[25] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[26] K. Park and T. Mulc, "Css10: A collection of single speaker speech datasets for 10 languages," *Interspeech*, 2019.

[27] J. M. K. Veaux, Christophe; Yamagishi, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit, [sound]," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017. [Online]. Available: https://doi.org/10.7488/ds/1994

[28] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[29] 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)*, 2018, pp. 5329–5333.

[30] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *35th International Conference on Machine Learning, ICML 2018*, vol. 12, pp. 8229–8238, 2018.

[31] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 10–21. [Online]. Available: https://aclanthology.org/K16-1002