The IOA-ThinkIT system for Blizzard Challenge 2021

Zengqiang Shang[1,2], *Ziyi Chen*[1,2], *Haozhe Zhang*[1,2], *Pengyuan Zhang*[1,2,*]

[1]Key Laboratory of Speech Acoustics Content Understanding, Institute of Acoustics, CAS, China
[2]University of Chinese Academy of Sciences, Beijing, China

zhangpengyuan@hccl.ioa.ac.cn

Abstract

In this paper, we introduce the bilingual text-to-speech system from IOA-ThinkIT to Blizzard Challenge 2021. This year's challenge aims to build a Spanish speech synthesis system, which also supports Spanish-English code-switch synthesis. We model the pronunciation, style and duration separately. For style modeling, Our approach adopts an analysis-synthesis scheme. At the analysis, a phoneme-level style encoder is utilized to extracted speaker-independent style vectors. Then an RNN auto-regressive predictor was built for style prediction at inference. We implement adversarial speaker training to text encoder of backbone and duration predictor to enable cross-language timbre transfer and cross-language duration transfer. Evaluation results provided by the challenge organizers are conducted over intelligibility, naturalness and similarity.

Index Terms: speech synthesis, Blizzard Challenge 2021, code-switch, bilingual TTS

1. Introduction

The Blizzard Challenge has been held once a year to evaluate different speech synthesis techniques based on same provided training database. It has made great contribution to the development of speech synthesis. This year Blizzard Challenge has two tasks:1. Hub task 2021-SH1 to synthesize European Spanish give about 5 hours speech data from a female native speaker of European Spanish; 2. Spoke task 2021-SS1 to synthesize the code-switch of European Spanish and English give the same dataset of task 1 while providing several natural recordings of such sentences as reference.

Speech synthesis aims to synthesis intelligible and natural speech from text, has a lot of applications in human communication and has long been a research topic in artificial intelligence, natural language and speech processing [1]. As the development of deep neural network, researchers are able to synthesis high-quality speech in recent years. For tradition methods like concatenating synthesis and statistic parametric speech synthesis (SPSS) [2], incorporating deep neural network in part of the TTS pipeline can significantly boost the performance in both naturalness and prosody. And in recent several years, speech synthesis systems based on deeper and larger neural network models have achieved near to human speech naturalness. In particular, auto-regressive models like Tacotron [3] [4] and Wavenet [5] take the advantage of the causal properties of speech, and demonstrate their strong performance. While traditional attention based acoustic end-to-end models are lack of robustness when generating speech, such as some of source text are repeated or skipped. So there are lots of work proposed for solving such problems, such as Duration Informed Attention Network and FastSpeech, which model the duration of phone directly to improve the stability and avoid these problems. Besides the base task of neural TTS, there are lots of other researches, such as emotional TTS, Style TTS, code-switch TTS, multilingual TTS and so on.

There are lots of exploring works on cross-lingual TTS. [6] presents a multi-lingual and multi-speaker neural TTS model based on VoiceLoop structure [7] with speaker and language embedding networks. [8] use a shared encoder with language embedding and two sperate language-dependent encoders Tacotron-based end-to-end systems by using a Mandarin and English monolingual speech data of two female speaker. [9] used a Tacotron2 based model to explore a Mandarin/English code-switched TTS model, which utilizes speaker embedding and phoneme-informed attention. [10] presents a multi-lingual TTS model, which trained on monolingual recordings from a large number of speakers. It uses a unified phoneme input representations and a adversarial loos to decouple speaker identities from speech content. [11] use cross-lingual voice conversion to get high quality data in other language, which help code-switched acoustics model train.

The paper is organized as follows. Section 2 introduce our system, including text analysis systems, acoustics model and vocoder. Section 3 presents the results of the benchmark systems and all the participation. Finally the conclusion is given in Section 4.

2. System Architecture

We follow the two-stage speech synthesis scheme and use Mel-spectrum as the intermediate representations. Our system consists of two parts, training and synthesis. At the training phase, we first prepossess the original data, and then train the spectrum model and vocoder. At the synthesis phase, we first do the text analysis which converts the test manuscripts to phoneme sequence and the utilize spectrum model and vocoder to get final waveform. We will introduce the details as follows.

2.1. Text analysis

The text analysis system mainly consists of graphame-to-phoneme (G2P) and language identification. We first perform language identification and then convert each language's text to phoneme sequence separately. For Spanish G2P conversion, we directly consult the pronunciation dictionary since the pronunciation is strictly corresponding to writing form in Spanish. And we utilize the open source [1] for English G2P conversion. For language identification, we use open tools [2] and set confidence 3 for each word.

We insert punctuation into phoneme sequence as input instead of use phoneme sequence only. Because punctuation

* is the corresponding author.

[1]https://github.com/Kyubyong/g2p
[2]https://github.com/facebookresearch/fastText

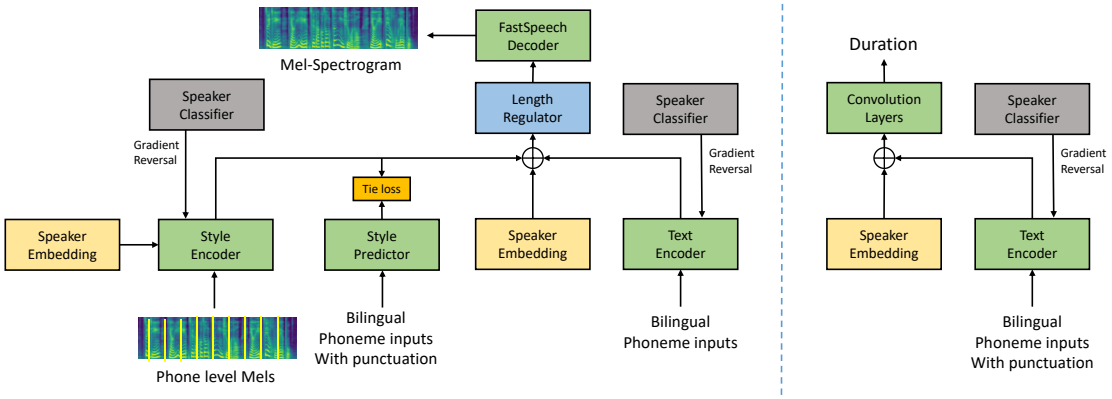

Figure 1: *The architecture of the spectrum model*

mask has a huge impact on intonation. For example, sentence end with "?" will have a obvious rising tone.

2.2. Spectrum Model

Follow our previous work [12], our spectrum model consists of three informational components. As show in Fig.1, which include fine-grained style encoder, bilingual text encoder, and speaker embedding, which separately contain dynamic style information speaker-independent linguistics information, and static timbre information. Moreover, a decoder component is added to fuse those information into Mel spectrum. Instead of training style predictor after style encoder converge, we train style encoder and style predictor simultaneously. And introducing a tie loss between encoded and predicted style vectors.

For text encoder, Each language has its phoneme set in our model, and all languages share the same self-attention based text encoder. To enable cross-language timbre cloning, we conducted an adversarial speaker classifier with a gradient reversal layer on the encoding representation to remove residual speaker information.

For style encoder, the fine-grained style encoder obtains temporal style representations from the aligned Mel-spectrograms at the phone-level, which maps each Mel-chunk into a fixed-length representation. We adopt a conditional variational encoding structure to extract speaker-independent style information, enabling cross-speaker style transfer without contaminating target timbre. We first apply three 2D convolution layers with batch norm for each segment. Then we add speaker embedding as conditional input, which is concatenated with the learned feature map and sent to a bi-directional Gated Recurrent Unit (GRU) that maps variable-length representation into fixed-length vector z_s . Then time-independent latent style representation z is obtained by passing z_s to a variational encoder, which is introduced to mitigate the sparseness of encoded representations. Furthermore, we constrain the dimension of latent space to 3 to function as an information bottleneck. We assume a prior distribution, $p(z) = N(z; 0, I)$, and train the model to maximize the evidence lower bound (ELBO) defined in Equation 1. We also add an adversarial speaker classifier with a gradient reversal layer over the mapped latent representation.

For style predictor, rather than using convolution layers describe in our previous work [12], we adopt a stack of feedforward transformer (FFT) blocks as the encoder of FastSpeech. Experiments show that predicted style vector is essential for naturalness. As show in 2, to improve the coherence of style prediction, we add a RNN autoregressive layer over FFT blocks. Which prove to be effective for improving the naturalness on both tasks. Unlike [13], which introduce a shallow autoregressive layer at the tail of frame decoder, we implement it in the phoneme encoder while keeping decoder parallel. Phoneme level auto-regression can improve performance without introducing too much computation than frame level one. Since our style encoder vectors are speaker-independent and language-dependent, style predictor takes phoneme punctuation sequence and output style vectors. We optimize the style predictor by minimize the mean square error. The gradient from tie loss also flow back to style encoder to extract speaker independent style vectors.

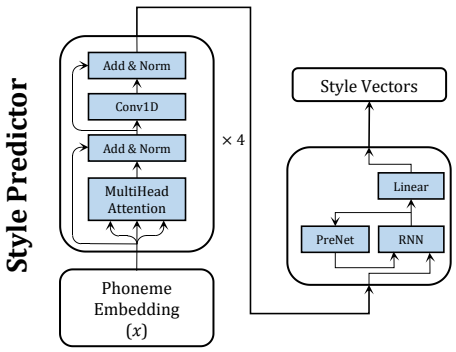

Figure 2: *The structure of the style predictor*

For duration predictor, we separately model the duration predictor instead of integrate it into pronunciation modeling [14]. We perform the speaker adversarial training using gradient reversal layer among duration predictor to cross language duration transfer. Experiment result suggests that it is important for a stable speaking speed especially in code-switching synthesis. Similarly, we optimize the duration predictor by minimizing the mean square error of duration.

For decoder, we follow the implement of FastSpeech [14]. And we expand it to multi-speaker version by introducing speaker embedding in a table lookup manner. We optimize the acoustic model by minimizing the mean square error of Mel spectrograms.

2.3. Neural Vocoder

we adopted 44.1k HIFIGAN for converting Mel spectrum to the waveform. We train the vocoder over given Spanish dataset. The code of HIFIGAN is from the official implementation [3]. We modified the transposed convolution's pad for the odd kernel size. For 44.1k HIFIGAN, we use upsample rates: [6, 5, 5, 4] and upsample kernel sizes: [12, 10, 10, 8], segment size is 12000, frame hop is 600 and frame length is 2400. And change resblock kernel size from [3, 7, 11] to [3, 5, 7, 11, 17].

2.4. Data processing

The training data provided by the organizer includes about 5 hours of speech data from a female native speaker of European Spanish. For English, we use our internal 44.1KHz dataset. Which is consists of 5 speaker and total duration is 5 hours. Firstly, the Montreal Force Aligner [15] is used to get the time-point of the initials and the finals. Secondly, we cut long sentence audio into pieces at the end of the short sentence in training and target set to ensure each clips' duration is less than 12.5 s. Finally, we converted the data sampling rate to 44.1kHz, trim the silence, and normalized audio loudness. Based on the processed data, we extract an 128-dimension Mel spectrum at 24kHz using a 50 ms frame length, 12.5 ms frame hop, and a Hann window function. Furthermore, we split out five samples from the target set as valid set.

3. Results

The listening test results in Blizzard Challenge 2021 were presented below. In this challenge, there are 13 (12 participating teams, plus natural speech) and 11 (10 participating teams, plus natural speech) systems in task 2021-SH1 and task 2021-SS1. Natural speech is marked as R, and our team identifier is I.

The evaluation comprised sections for task SH1 and task SS1. Task SH1 evaluates three aspects, including naturalness, similarity and intelligibility, while task SS1 evaluates naturalness, similarity and acceptability of English words.

3.1. Naturalness evaluation

In naturalness evaluating section, each audio sample was evaluated over Mean Opinion Score (MOS). Listeners listened to one sample and chose a score which represented how natural or unnatural the sentence sounded on a scale of 1 [Completely Unnatural] to 5 [Completely Natural]. Figure 4 show the boxplot of all systems' evaluation results on naturalness in task SH1 and task SS1. Our system achieves an average score of 3.55 and 3.43 in task SH1 and task SS1.

We trained the acoustics model and vocoder separately, which may cause a mismatch in the system and influence the naturalness of synthesized speech. After listening to submitted audios, some examples have low speech quality even if the over performance is acceptable as shown in 3. For code-switch task SS1, separated modeling pronunciation, duration, and style are essential for natural prosody. Speaker adversarial training over three components enables cross-language timbre transfer, cross-language duration transfer, and cross-language style transfer, respectively.

[3]https://github.com/jik876/hifi-gan

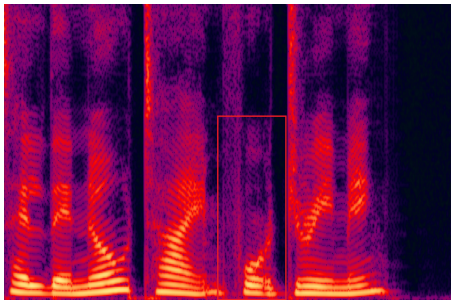

Figure 3: *MEL spectrogarm generated from our system.*

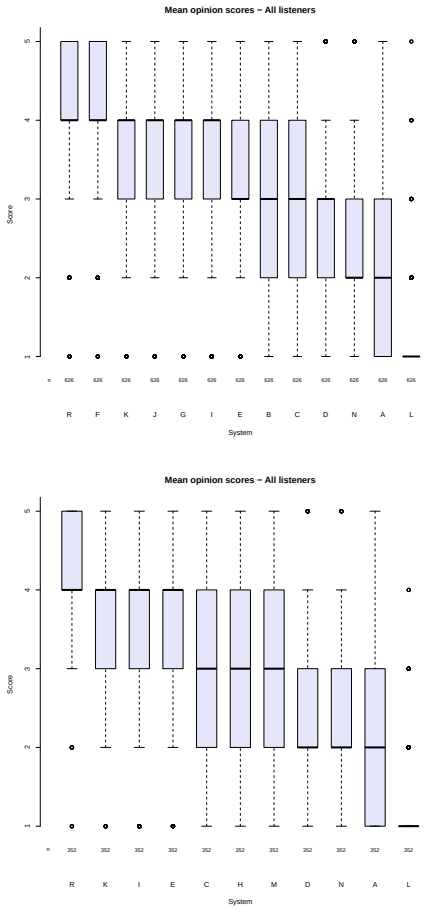

Figure 4: *Boxplot of MOS.*

3.2. Similarity evaluation

Figure 5 shows the boxplot of evaluation results of all systems on Spanish similarity. Our system I achieved the MOS of 3.89 and 3.76 in task SH1 and task SS1. We used only MSE loss on mel-spectrograms to train the spectrum model, so the model may lose some detailed information influencing speaker similarity. However, Speaker adversarial training over text encoder is vital to enable cross-language timbre transfer in SS1.

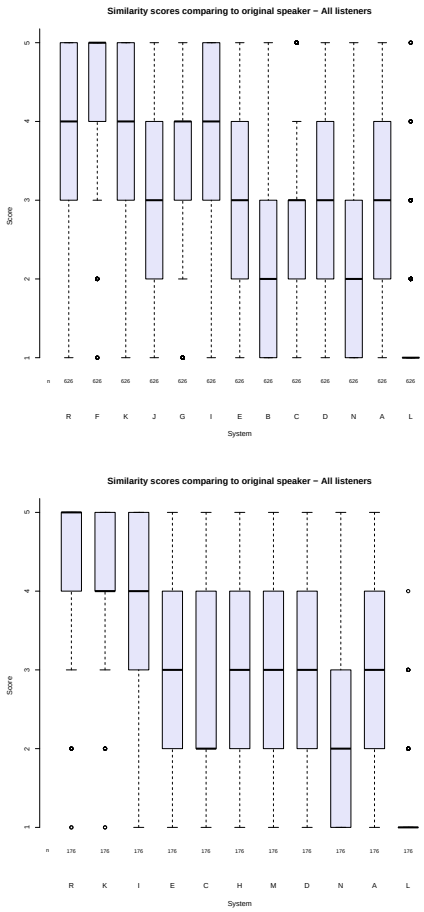

Figure 5: *Boxplot of SIM.*

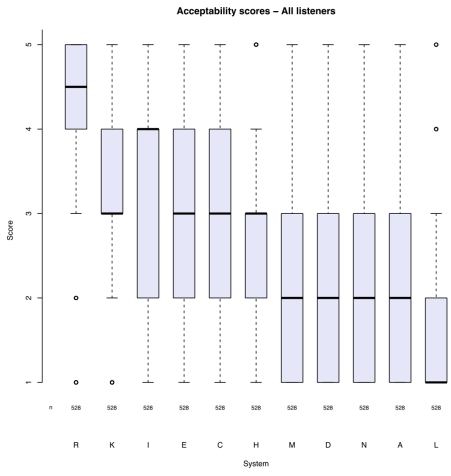

Figure 6: *Boxplot of ACC.*

3.3. Acceptability evaluation

Figure 6 shows the boxplot of evaluation results of all systems on acceptability of English words in task SS1. By conducting an adversarial speaker classifier with a gradient reversal layer on the text encoder in acoustics model, our system can decouple the residual speaker information and text information, which improved the acceptability of English words. It is hard to distinguish English from Spanish text, which also affects the acceptability of English words.

3.4. Intelligibility evaluation

The boxplot of PER results for Sharvard test and SUS test in task SS1 were presented in Figure.7.8. Our team has no native Spanish speaker, so it is hard to find mistakes in our front-end models. We perform G2P for Spanish by consulting pronunciation dictionary download from [15]. It way cause some errors. We focus more on speaker similarity and speech quality.

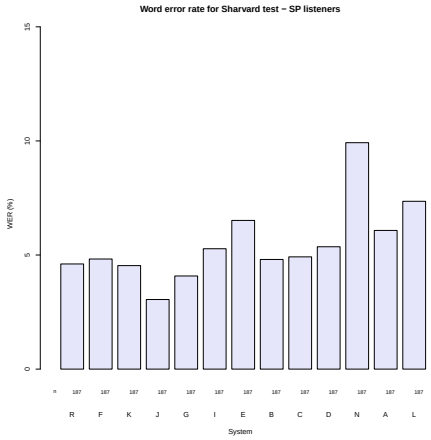

Figure 7: *Boxplot of INT SHARVARD.*

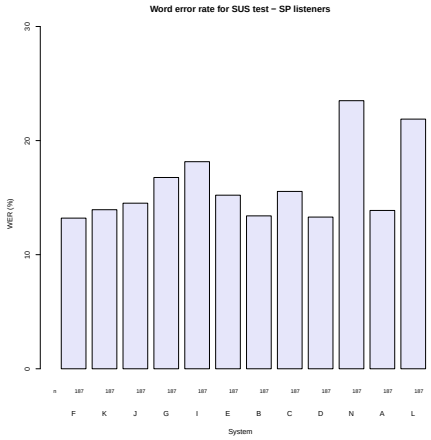

Figure 8: *Boxplot of INT SUS.*

4. Conclusions

This paper presents our IOA-ThinkIT system submitted for Blizzard Challenge 2021. We focus on multilingual spectrum modeling and model pronunciation, style and duration separately. The effectiveness of our system on code-switch synthesis is successfully confirmed by the official evaluation results.

5. References

[1] R. B. Adler and G. R. Rodman, *Understanding human communication*, 2016, vol. 13.

[2] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 ieee international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7962–7966.

[3] 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 *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. ISCA, 2017.

[4] 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.

[5] 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," *SSW*, p. 125, 2016.

[6] E. Nachmani and L. Wolf, "Unsupervised polyglot text-to-speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7055–7059.

[7] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," in *International Conference on Learning Representations*, 2018.

[8] Y. Cao, X. Wu, S. Liu, J. Yu, X. Li, Z. Wu, X. Liu, and H. Meng, "End-to-end code-switched tts with mix of monolingual recordings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6935–6939.

[9] L. Xue, W. Song, G. Xu, L. Xie, and Z. Wu, "Building a mixed-lingual neural tts system with only monolingual data," 2019.

[10] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. 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," 2019.

[11] 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," 2020.

[12] Z. Shang, Z. Huang, H. Zhang, P. Zhang, and Y. Yan, "Incorporating cross-speaker style transfer for multi-language text-to-speech," *Proc. Interspeech 2021*, pp. 1619–1623, 2021.

[13] R. Liu, B. Sisman, Y. Lin, and H. Li, "Fasttalker: A neural text-to-speech architecture with shallow and group autoregression," *Neural Networks*, vol. 141, pp. 306–314, 2021.

[14] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: fast, robust and controllable text to speech," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 3171–3180.

[15] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.