Interspeech 2022
18-22 September 2022, Incheon, Korea

SANE-TTS: Stable And Natural End-to-End Multilingual Text-to-Speech

Hyunjae Cho[1,2], *Wonbin Jung*[1,3], *Junhyeok Lee*[1], *Sang Hoon Woo*[1]

[1]MINDsLab Inc., Republic of Korea
[2]Seoul National University, Republic of Korea
[3]Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
{chohyunjae, wbjung, jun3518, shwoo}@mindslab.ai

Abstract

In this paper, we present SANE-TTS, a stable and natural end-to-end multilingual TTS model. By the difficulty of obtaining multilingual corpus for given speaker, training multilingual TTS model with monolingual corpora is unavoidable. We introduce speaker regularization loss that improves speech naturalness during cross-lingual synthesis as well as domain adversarial training, which is applied in other multilingual TTS models. Furthermore, by adding speaker regularization loss, replacing speaker embedding with zero vector in duration predictor stabilizes cross-lingual inference. With this replacement, our model generates speeches with moderate rhythm regardless of source speaker in cross-lingual synthesis. In MOS evaluation, SANE-TTS achieves naturalness score above 3.80 both in cross-lingual and intralingual synthesis, where the ground truth score is 3.99. Also, SANE-TTS maintains speaker similarity close to that of ground truth even in cross-lingual inference. Audio samples are available on our web page[1].

Index Terms: multilingual text-to-speech, cross-lingual speech synthesis, text-to-speech, domain adversarial training

1. Introduction

While almost all works in text-to-speech (TTS) focus on synthesizing speech in a single language, multilingual TTS aims to generate speeches in multiple languages with a single model. The most naive approach is training a model with multilingual speech dataset for a given speaker, but it is infeasible due to the unavailability of such datasets. Thus, prior works [1, 2, 3, 4, 5] have focused on using a mix of monolingual corpora to build multilingual TTS models by implementing cross-lingual speech synthesis.

Previous multilingual TTS models [1, 2, 3, 4, 5] are mainly based on Tacotron [6, 7]. However, Tacotron-based models have several issues while they synthesize speeches in an autoregressive manner that utilizes attention [8] to align input text and target speech. First, attention errors cause wrong alignment estimation, resulting in the problem of words skipping and repeating [9]. Second, generating results autoregressively with attention inhibits direct control of phoneme-level duration [10]. On the other hand, some multilingual TTS models [11, 12] are not Tacotron-based models. YourTTS [11] focuses more on zero-shot learning and does not support cross-lingual synthesis. Wu et al. [12] also proposed a multilingual TTS model based on voice conversion. However, this study is closer to TTS data augmentation by the pre-trained voice conversion model and only covers the Indo-European languages with the International Phonetic Alphabet representation.

[1]https://mindslab-ai.github.io/sane-tts/

Most non-autoregressive models [10, 13, 14, 15, 16, 17] include duration predictor to estimate each phoneme's duration and total length without autoregressive iteration. Among the duration predictor-based models [10, 13, 14, 15, 16, 17, 18, 19], we choose VITS [17], an end-to-end TTS model, as a backbone. Furthermore, VITS is advantageous because it generates natural speeches with high synthesis speed and requires a single model training since it is not a two-stage model. While utilizing a duration predictor, the key challenge for multilingual TTS is the uncertainty of duration prediction in cross-lingual synthesis.

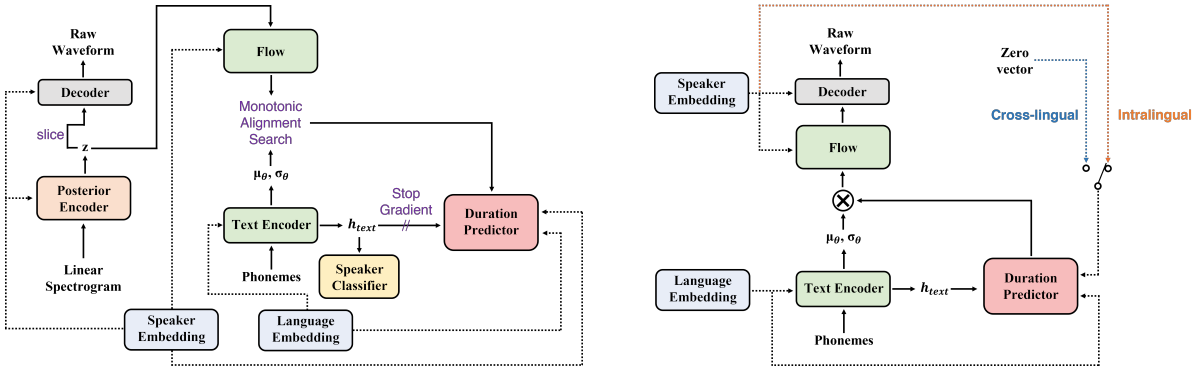
In this paper, we propose *SANE-TTS*, a multilingual TTS model with natural speech synthesis and stable duration prediction. Due to using monolingual corpora, speaker identity and linguistic features can be entangled. To generate natural speeches in cross-lingual synthesis, speaker identity needs to disentangle from linguistic features. So, we add *speaker regularization loss* term to prevent language information leakage to speaker representation. To predict duration stably, duration predictor should produce phoneme duration independent of speaker identity during cross-lingual inference. For this reason, our duration predictor uses zero vector instead of speaker embedding generating moderate phoneme duration regardless of speaker identity. Our contributions can be summarized as:

- *SANE-TTS* synthesizes multilingual speeches stably with a perceptual score close to the ground truth level while maintaining speaker similarity even in cross-lingual inference.
- Proposed *speaker regularization loss* achieves improvement of speech naturalness as much as domain adversarial training (DAT) [20], which is commonly used in previous cross-lingual synthesis studies [2, 3].

2. Method

We modify some modules and loss terms to build multilingual TTS model. To receive various languages, we change text encoder and duration predictor. For loss function, we apply DAT [20] to make text representation disentangle from speaker identity. Also, we add a speaker regularization loss term to learn language-independent speaker representations. Figure 1 illustrates overview of our system during training procedure and inference procedure. SANE-TTS gets phoneme sequences, speaker embedding, and language embedding, as inputs, and generates raw waveform as an output. We use different phoneme sets and grapheme-to-phoneme converters for each language during conversion of transcripts into phoneme sequences. In training procedure, posterior encoder gets linear spectrogram as an additional input.

10.21437/Interspeech.2022-46



(a) *Training procedure* (b) *Inference procedure*

Figure 1: *Block diagram of system overview of proposed model in (a) training procedure and (b) inference procedure.*

2.1. Language embedding

SANE-TTS includes learnable language embeddings, which are 256-dimensional vectors, the same size with the speaker embeddings. We only provide language embedding to text encoder and duration predictor, which directly get text representation as an input. To provide language information in duration predictor, language embedding is passed through a convolution layer and added to the hidden text representation in the same manner as speaker embedding. For the flow, the posterior encoder, and the decoder, we retain the same setup as the original VITS.

2.2. Text encoder

We use a Transformer-based text encoder [21] with relative positional representation [22] same as VITS, with parameter generation suggested by Nekvinda and Dušek [2] to be compatible with various languages. The parameter generator takes language embedding as an input and generates parameters of the text encoder. This process helps optimizing the text encoder for each language of input phoneme sequences.

2.3. Domain adversarial training

Since every speaker does not share the same transcript, text representation in TTS model can be entangled with speaker identity. To reduce speakers' bias on the text representation, we use DAT [3, 20]. We attach a speaker classifier at the end of the text encoder as a domain classifier. The speaker classifier consists of fully connected layers, and a gradient reversal layer is inserted between the speaker classifier and the text encoder. We train the speaker classifier with cross-entropy loss to prevent predicting speaker identity from text representation. Through DAT, the text encoder learns speaker-independent text representation, and the model can generate speeches from general texts.

2.4. Speaker regularization loss

Similar to speaker's bias on the text representation, speaker identity is biased by the language of the speaker's utterances in the dataset. To synthesize speech across various languages, we have to prevent speaker identities from entangling to languages in the duration predictor. To reduce speaker bias for language, we introduce a *speaker regularization loss* L_{reg} given by:

$$L_{\text{reg}} = \| \mathbb{E}_{k \in K} \left[\text{conv}(S_k) \right] \|_2, \tag{1}$$

where conv is a convolution layer with a kernel size of 1 and S_k represents the speaker embedding of a speaker in datapoint k in the batch K. As the mean of hidden speaker representations $\text{conv}(S_k)$ is pushed to zero vector regardless of the language, the speaker identities disentangle to languages in the model.

Since the proposed speaker regularization loss pushes the mean of hidden speaker representations toward zero vector, the duration predictor estimates moderate phoneme duration by inputting a zero vector instead of speaker embedding in cross-lingual inference. On the other hand, the duration predictor gets speaker embedding as an input during intralingual inference because input text consists of seen phonemes by the speaker. This method reduces the instability of the duration predictor and removes the uncertainty of adjusting the speaker information to phoneme duration in cross-lingual synthesis.

2.5. Deterministic duration predictor

Originally in VITS, a stochastic duration predictor (SDP) was proposed to generate speeches with diverse rhythms. SDP predicts phoneme duration stochastically from noise latent by normalizing flow [23]. However, Casanova et al. [11] reported that there are cases where the SDP generates unnatural duration causing unclear pronunciation. So, in this paper, a deterministic duration predictor (DDP) [13, 17] is applied to improve the stability of speech synthesis.

3. Experiments

3.1. Dataset

We construct our dataset by gathering internal and external speech corpora [24, 25, 26, 27, 28] in four languages. The dataset is composed of speeches from multiple speakers in English (EN), Korean (KO), Japanese (JA) and Mandarin Chinese (ZH) as shown in Table 1. We resample audio samples to 22.05 kHz. We convert transcripts of speeches into phoneme sequences through our internal grapheme-to-phoneme conversion process. We hold out 5% of utterances for validation set.

Table 1: *Details of our dataset in multiple languages*

Language	EN	KO	JA	ZH	Total
Number of speakers	161	27	110	174	472
Length	72.4hr	43.5hr	27.5hr	60.0hr	203.4 hr

3.2. Setup

We train our model for 200 epochs on 2 NVIDIA A100 GPUs. We use mixed precision training with batch size of 64. We follow the schedule of scale factor λ which weights the speaker classification loss in DAT [20] as:

$$\lambda = \frac{2}{1 + \exp(-10 \cdot p)} - 1, \tag{2}$$

where p is the training progress linearly changing with the training steps, from 0 to 1. The loss scale factor is initiated at 0 and continues to increase as training progresses. This schedule suppresses DAT at the early stage and improves the quality of outputs. Other details follow that of VITS [17].

3.3. Baseline model

To demonstrate that our model generates high-quality speeches during cross-lingual inference, we compare SANE-TTS with an official open-sourced implementation of another multilingual TTS model (meta-learning model)[2] proposed by Nekvinda and Dušek [2]. We train meta-learning model up to 100 epochs with a batch size of 64. For vocoding, we use the Griffin-Lim algorithm [29] provided in the official implementation.

3.4. Ablation study

We compare SANE-TTS with models that remove a single modification from our model. We remove three modifications; applying DAT, adding speaker regularization loss, and replacing SDP by DDP. When removing our speaker regularization loss, we input speaker embedding in the duration predictor during cross-lingual inference. We exclude modifications regarding language embedding, which is necessary to implement our multilingual TTS model.

3.5. Evaluation

We report the mean opinion scores (MOS) to evaluate the naturalness of speeches and the similarity of the speaker between speech pairs of ground truth and synthesized samples [30], including 95% confidence intervals. Our MOS evaluation uses an absolute category rating scale, where raters score performances from 1 to 5 in 1 point increments. We conduct MOS evaluation on Amazon Mechanical Turk framework. Due to the difficulty of acquiring native speakers for rating non-English target languages, we gather opinion scores on English speeches from native English speakers. Each speech and speech pair is scored by 5 raters. For speaker similarity, raters compare a synthesized speech with a speech of the same speaker in the validation set and evaluate speaker similarity between them. In the case of ground truth, we select two speeches of the same speaker in the validation set and raters evaluate their speaker similarity.

To generate evaluation speech samples, we sample 30 sentences randomly from the English validation set. Also, we select 5 male and 5 female speakers from every four languages to evaluate both the intralingual and the cross-lingual synthesis capabilities. We synthesize speeches for every combination of speakers and sentences, total 1200 utterances.

[2] https://github.com/Tomiinek/Multilingual_Text_to_Speech

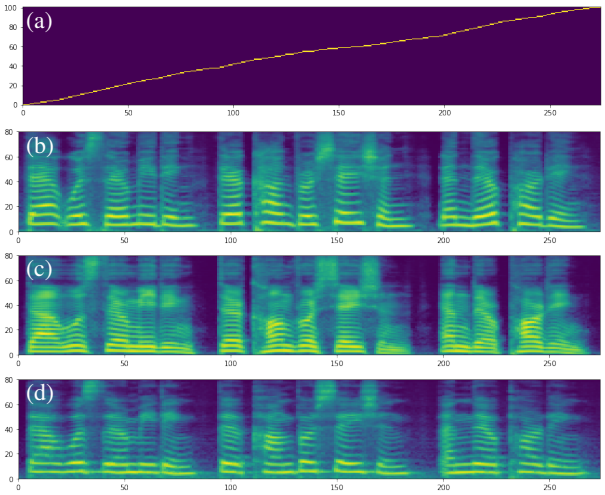


Figure 2: *Cross-lingual alignment and mel spectrograms of the text "That was their meeting, their conversation, and their parting." (a) is cross-lingual alignment between the text and speech. (b-d) are from synthesized speech of (b) KSS (KO) [26], (c) jvs001 (JA) [27], and (d) SSB0018 (ZH) [28].*

4. Results and Discussion

4.1. Speech synthesis quality

Table 2 and Table 3 show naturalness MOS and similarity MOS of the comparative models with 95% confidence intervals. SANE-TTS surpasses meta-learning model both in naturalness and speaker similarity. Our model achieves naturalness MOS of 3.95 for the intralingual synthesis which is close to that of the ground truth 3.99. Also, naturalness of the cross-lingual synthesis does not change significantly in the intralingual synthesis, above 3.80 MOS. For both naturalness MOS and similarity MOS, English speech samples get higher scores than other languages. It is because of different levels of difficulty of intralingual and cross-lingual synthesis. Also, the evaluation could be biased by the raters who are English speakers.

Table 2: *Comparison of naturalness MOS with another model*

Model	Intralingual	Cross-lingual		
	EN	KO	JA	ZH
Ground truth	3.99 ± 0.04	-	-	-
SANE-TTS	$\mathbf{3.95 \pm 0.04}$	$\mathbf{3.80 \pm 0.05}$	$\mathbf{3.84 \pm 0.04}$	$\mathbf{3.81 \pm 0.04}$
Meta-learning model	3.19 ± 0.06	3.30 ± 0.06	3.03 ± 0.06	2.97 ± 0.06

Table 3: *Comparison of similarity MOS with another model*

Model	Intralingual	Cross-lingual		
	EN	KO	JA	ZH
Ground truth	3.38 ± 0.05	3.60 ± 0.05	3.44 ± 0.05	3.50 ± 0.05
SANE-TTS	$\mathbf{3.48 \pm 0.06}$	$\mathbf{3.31 \pm 0.06}$	$\mathbf{3.26 \pm 0.06}$	$\mathbf{3.44 \pm 0.06}$
Meta-learning model	2.92 ± 0.06	2.81 ± 0.07	2.59 ± 0.06	2.73 ± 0.06

4.1.1. The cross-lingual speech synthesis

Figure 2 shows an alignment and mel spectrograms of speeches synthesized by SANE-TTS during cross-lingual inference. Our

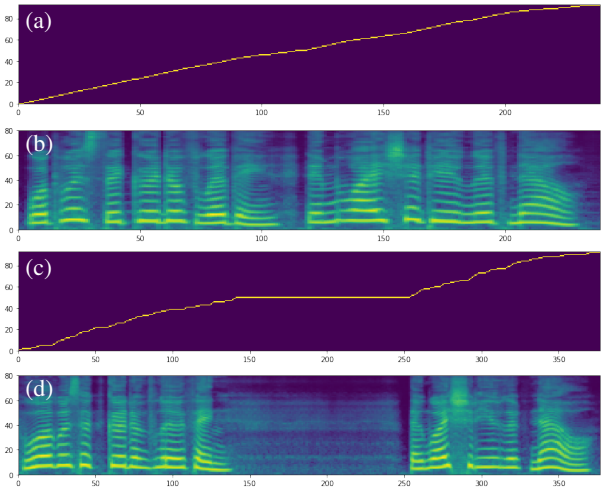


Figure 3: *Alignments and mel spectrograms of the speech generated by models with DDP and SDP. The text of the speech is "What was the good of living, and why should he live now?" and the speaker is KSS [26]. The model with DDP generates (a) alignment and (b) mel spectrogram. In extreme cases, model with SDP generates (c) alignment and (d) mel spectrogram with overlong silence.*

model generates identical alignment for every speaker since DDP receives a zero vector instead of speaker embedding. Thus, SANE-TTS synthesizes speeches with moderate rhythm regardless of source speaker.

4.2. Ablation study

Table 4 and Table 5 show naturalness MOS and similarity MOS of the ablation study with 95% confidence intervals. SANE-TTS shows degradation of naturalness MOS in the cross-lingual inference without the proposed speaker regularization loss upto 0.11 and DAT upto 0.06. Specifically, removing speaker regularization loss decreases speech naturalness slightly more than excluding DAT in cross-lingual synthesis.

There is no statistically significant difference between the ablation model with SDP and our model. However, there are cases where SDP predicts an unnatural duration in cross-lingual synthesis. Figure 3 depicts generated alignments and mel spectrograms from models with DDP and SDP. The SDP predicts unnatural duration that produces an overlong silence in the middle of the speech. Although SDP produces speeches with diverse rhythms, we apply relatively stable DDP for reliable multilingual TTS model.

Table 4: *Comparison of naturalness MOS in the ablation study*

Model	Intralingual	Cross-lingual		
	EN	KO	JA	ZH
Ground truth	3.99 ± 0.04	-	-	-
SANE-TTS	$\mathbf{3.95 \pm 0.04}$	$\mathbf{3.80 \pm 0.05}$	$\mathbf{3.84 \pm 0.04}$	$\mathbf{3.81 \pm 0.04}$
w/o DAT	3.85 ± 0.05	3.77 ± 0.04	3.82 ± 0.04	3.75 ± 0.04
w/o Regularization	3.88 ± 0.04	3.69 ± 0.05	3.76 ± 0.04	3.72 ± 0.05
w/ SDP	3.93 ± 0.04	$\mathbf{3.81 \pm 0.05}$	3.80 ± 0.04	3.74 ± 0.05

Table 5: *Comparison of similarity MOS in the ablation study*

Model	Intralingual	Cross-lingual		
	EN	KO	JA	ZH
Ground truth	3.38 ± 0.05	3.60 ± 0.05	3.44 ± 0.05	3.50 ± 0.05
SANE-TTS	$\mathbf{3.48 \pm 0.06}$	$\mathbf{3.31 \pm 0.06}$	$\mathbf{3.26 \pm 0.06}$	$\mathbf{3.44 \pm 0.06}$
w/o DAT	3.33 ± 0.06	3.16 ± 0.06	3.15 ± 0.06	3.40 ± 0.06
w/o Regularization	3.34 ± 0.06	3.29 ± 0.06	3.18 ± 0.06	3.47 ± 0.06
w/ SDP	3.45 ± 0.06	$\mathbf{3.34 \pm 0.06}$	3.26 ± 0.06	$\mathbf{3.48 \pm 0.06}$

4.2.1. Visualizing regularization of speaker identities

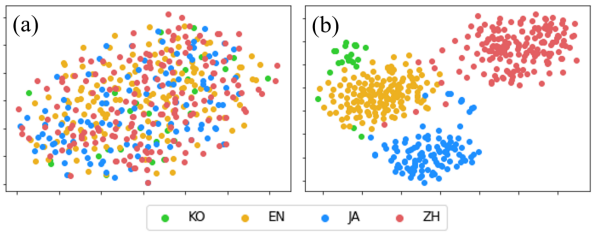


Figure 4: *t-SNE plots of hidden speaker representations for the duration predictor for every 472 speakers (a) with speaker regularization loss and (b) without speaker regularization loss.*

Figure 4 shows the t-SNE plots of hidden speaker representations for the duration predictor with and without the speaker regularization loss. Without the speaker regularization loss, the hidden speaker representations form clusters by languages, while no clusters are observed, and samples are distributed around the center with the speaker regularization loss. The t-SNE plots demonstrate that our model learns language-independent speaker representations by adding the speaker regularization loss.

5. Conclusions

In this paper, we propose SANE-TTS, a stable and natural end-to-end multilingual TTS model. Due to the limited multilingual speech corpus, we use a mix of monolingual corpora for training multilingual TTS model. Therefore, multilingual TTS model faces the difficulty of cross-lingual inference for languages that are not recorded by target speaker. To solve difficulty of multilingual TTS, we introduce *speaker regularization loss* to make our model learns speaker representation independently from its own language. Also, by replacing speaker embedding with zero vector in the cross-lingual duration prediction, the model produces moderate phoneme duration irrelevant to speaker identity. In addition, we add language embedding and apply DAT which are commonly used techniques. In our multilingual setup with English, Korean, Japanese and Mandarin Chinese, SANE-TTS generates natural audio samples which obtain high speaker similarity during both the cross-lingual and the intralingual synthesis. Based on our study, we expect to expand SANE-TTS into other languages in future work.

6. Acknowledgements

The authors would like to thank Seungu Han and Kangwook Kim from MINDsLab Inc., Jinwoo Kim from KAIST, and Hyeongkeun Kim from University of Illinois Urbana-Champaign for providing beneficial feedback on this work.

7. References

[1] 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 *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6935–6939.

[2] T. Nekvinda and O. Dušek, "One Model, Many Languages: Meta-Learning for Multilingual Text-to-Speech," in *INTERSPEECH*, 2020, pp. 2972–2976.

[3] 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," in *INTERSPEECH*, 2019, pp. 2080–2084.

[4] Z. Liu and B. Mak, "Cross-lingual Multi-speaker Text-to-speech Synthesis for Voice Cloning without Using Parallel Corpus for Unseen Speakers," *arXiv preprint arXiv:1911.11601*, 2019.

[5] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, and J. Xiao, "Cross-Lingual, Multi-Speaker Text-To-Speech Synthesis Using Neural Speaker Embedding," in *INTERSPEECH*, 2019, pp. 2105–2109.

[6] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," in *INTERSPEECH*, 2017, pp. 4006–4010.

[7] 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," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *International Conference on Learning Representations*, 2015.

[9] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: 2000-Speaker Neural Text-to-Speech," in *International Conference on Learning Representations*, 2018.

[10] 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 *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[11] E. Casanova, J. Weber, C. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone," *arXiv preprint arXiv:2112.02418*, 2021.

[12] J. Wu, A. Polyak, Y. Taigman, J. Fong, P. Agrawal, and Q. He, "Multilingual Text-To-Speech Training Using Cross Language Voice Conversion And Self-Supervised Learning Of Speech Representations," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8017–8021.

[13] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," in *Advances in Neural Information Processing Systems*, 2020.

[14] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *International Conference on Learning Representations*, 2021.

[15] A. Łańcucki, "Fastpitch: Parallel Text-to-Speech with Pitch Prediction," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.

[16] T. Bak, J.-S. Bae, H. Bae, Y.-I. Kim, and H.-Y. Cho, "FastPitchFormant: Source-Filter Based Decomposed Modeling for Speech Synthesis," in *INTERSPEECH*, 2021, pp. 116–120.

[17] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *International Conference on Machine Learning*, vol. 139, 2021, pp. 5530–5540.

[18] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-StyleSpeech : Multi-Speaker Adaptive Text-to-Speech Generation," in *International Conference on Machine Learning*, vol. 139, 2021, pp. 7748–7759.

[19] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. Skerry-Ryan, and Y. Wu, "Parallel Tacotron 2: A Non-Autoregressive Neural TTS Model with Differentiable Duration Modeling," in *INTERSPEECH*, 2021, pp. 141–145.

[20] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[22] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-Attention with Relative Position Representations," in *NAACL-HLT (2)*, 2018, pp. 464–468.

[23] D. J. Rezende and S. Mohamed, "Variational Inference with Normalizing Flows," in *International Conference on Machine Learning*, 2015.

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

[25] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *INTERSPEECH*, 2019, pp. 1526–1530.

[26] K. Park, "KSS Dataset: Korean Single speaker Speech Dataset," https://kaggle.com/bryanpark/korean-single-speaker-speech-dataset, 2018.

[27] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.

[28] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A Multi-speaker Mandarin TTS Corpus and the Baselines," *arXiv preprint arXiv:2010.11567*, 2015.

[29] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[30] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno, and Y. Wu, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.