INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

STEN-TTS: Improving Zero-shot Cross-Lingual Transfer for Multi-Lingual TTS with Style-Enhanced Normalization Diffusion Framework

Chung Tran[1], *Chi Mai Luong*[2], *Sakriani Sakti*[1]

[1]Japan Advanced Institute of Science and Technology (JAIST), Japan
[2]Institute of Information Technology (IOIT), Vietnam

chungtq@jaist.ac.jp, lcmai@ioit.ac.vn, ssakti@jaist.ac.jp

Abstract

The prevalence of personalized multilingual tools plays an important role in learning aids and virtual assistants. The existing works on multilingual adaptive text-to-speech (TTS) mainly focus on fine-tuning models or extracting personal styles, such as prosody, emotion, and identity, with the aim of adapting to new speakers. This paper introduces the Style-Enhanced Normalization TTS (STEN-TTS) approach to synthesizing multilingual voice and maintaining personal styles with only 3 seconds of input reference. By presenting an integrated module (STEN) into the diffusion model, the proposed method can simulate the speaker's style and eliminate white noise in the synthesized speech. The experimental results show that our model achieves good performance, at above 3.5 on SMOS for cross-lingual switching. Furthermore, when using speaker verification to assess the similarity between the ground truth and synthesized voices, the accuracy reaches 82.4% with 3 seconds of audio reference.

Index Terms: multilingual text-to-speech, adaptation, diffusion model

1. Introduction

In recent years, the demand for multilingual adaptive TTS has grown rapidly in modern society. This technology has a wide range of significant uses, including tailored language learning programs and virtual assistants that can interact with users in their preferred language and voice. Developing a system that can accurately copy a user's speaking characteristics and translate them into a different language is challenging because it is impractical and expensive to collect the relevant data from speakers of multiple languages. In addition, it is not easy to copy individualized speech because the human voice contains a large amount of information, including identity, prosody, and emotion. For this reason, the study of multilingual adaptive TTS poses difficulties for speech-processing experts.

Several techniques [1, 2] have attempted to synthesize speech by fine-tuning a model using a sample size of 1-5 minutes for adaptation. Nonetheless, these techniques optimize output through several iterations or epochs, requiring audio and corresponding transcripts. In addition, a speaker embedding module [3, 4, 5] is used by other methods to extract a latent vector from the audio reference. This speaker vector can capture crucial details of personal styles, including prosody and emotion. Frameworks, such as Meta-StyleSpeech [3] and AdaSpeech4 [6], have demonstrated their ability for high adaptation with the short audio input, of around a few seconds, particularly for the unseen speakers. Nevertheless, due to language features and speaking style variations, the speaker output differs from the input when using these approaches to synthesize cross-lingual speech.

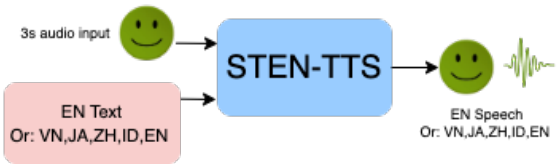

Figure 1: *Overview of STEN-TTS*

For the multilingual TTS domain, many studies have applied Tacotron2 [7, 8, 9], since its outstanding output synthesis. However, the time needed for inference in this autoregressive model is significantly long, especially for lengthy phrases or paragraphs and for certain words that are frequently repeated or skipped. Recently, several speech researchers have exploited the potential of the non-autoregressive models due to its robustness and inference speed. When training a singer speaker, the capabilities of these models, namely FastSpeech2 [10] and VITS [11], make it possible to synthesize high-fidelity audio. However, training with multiple speakers produces white noise in synthesized speech. Models, such as Your-TTS [12] and SANE-TTS [13], build on VITS [11], produce output audio degraded by white noise when synthesizing cross-lingual speech with very short audio input.

Lately, a diffusion probabilistic model [14] has been proven capable of synthesizing high-quality images and eliminating white noise in speech processing. More specifically, the diffusion model consists of two phases, which are the forward and reverse processes. In the forward process, the model adds a small noise to the input and turns it into the gaussian distribution noise. In the reverse process, the denoising model tries to disentangle the input from the gaussian noise and recover the original data in K time steps. Methods, such as DiffSinger [15], and DiffGan-TTS [16], have achieved good results by applying diffusion models to TTS. The synthesized speech using these models has high-fidelity, naturalness, and non-white noise. However, most studies exploit diffusion on a single speaker, and output voices lose personal styles when applied to multiple speakers.

Hence, as a way to overcome the weaknesses of this diffusion model, we propose Style-Enhanced Normalization text-to-speech (STEN-TTS), as shown in Figure 1, to take advantage of the diffusion model and solve the problem of losing personal style information during the reverse process by STEN. Our experiments demonstrate the module's effectiveness as shown by the SMOS score. Our main contributions are as follows:

- To our best knowledge, this is the first diffusion framework that allows adapting by cross-lingual synthesis with only 3

10.21437/Interspeech.2023-2243

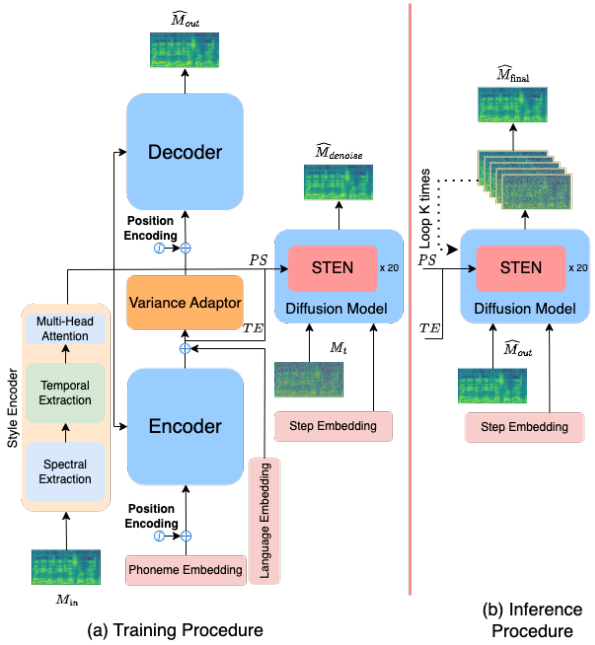

Figure 2: *Proposed architecture of STEN-TTS. Figure (a) describes the training process. Figure (b) shows the inference procedure with K time steps.*

seconds of audio input.

- Second, we introduce the Style-Enhanced Normalization (STEN) module to the diffusion model, which can enhance the personal styles of both seen and unseen speakers.
- Third, our proposed model achieves good results in terms of SMOS score for cross-lingual adaptation in five languages: English, Chinese, Japanese, Indonesian, and Vietnamese.

2. Method

2.1. Text-to-Speech

The overall architecture of the TTS model built on FastSpeech2 comprises the encoder, variance adaptor, and decoder as shown in Figure 2-a. The encoder takes the phoneme embedding as the input and converts it into a hidden sequence before combining it with the personal style (PS) vector by the Style-Adaptive Layer Norm [3]. Next, at the variance adaptor, this hidden sequence is augmented with other information, namely duration for each phoneme, pitch contour, and energy. Similarly to the encoder, the decoder fuses this hidden sequence and PS vector to produce the final mel-spectrogram, which can be converted to a signal wave using vocoders such as Hifi-GAN [17].

2.2. Style Encoder

A few seconds of audio X is converted to mel-spectrogram M_{in} before being input to the Style Encoder (StyEnc) [3] to extract the personal style (PS), as shown in Figure 2-a. The output of this module is a 128-dimensional vector. More specifically, the StyEnc module has three blocks, each with its own functionality: spectral extraction, temporal extraction, and multi-head attention.

$$PS = \text{StyEnc}(M_{in}), PS \in \mathbb{R}^N. \quad (1)$$

- **Spectral Extraction**: This module has three linear layers, followed by spectral normalization and the Mish activation function. This block receives mel-spectrogram (M_{in}) as the input and transforms it into a sequence of the feature vector.
- **Temporal Extraction**: This module includes, consecutively, convolutions and batch normalization, which are used to capture all important information of the input speaker.
- **Multi-head Attention**: Next, as the output of the temporal block passes to Multi-head Attention, this module obtains the principal features of the speaker by applying three operations: query, key, and value. These features are passed to the fully connected layer, where temporal average pooling is used to compress them into a 128-dimensional vector.

2.3. Style-Enhanced Diffusion Mechanism

The Diffusion model [14] adds gaussian noise to the initial data sample during the training process as shown in Figure 2-a. After T time steps, the model turns the input data into a gaussian noise distribution:

$$M_t = \sqrt{\bar{\alpha}_t}(M_{t-1}) + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t \in (1, ..., T), \quad (2)$$

where the input sample $M_0 = M_{in}$, and after adding small noises consecutively in T time steps, this end with $M_T \sim \mathcal{N}(0, \mathbf{I})$. Furthermore $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Then by applying the Markov chain, we have the following formula:

$$q(M_t|M_0) = \mathcal{N}(M_t; \sqrt{\bar{\alpha}_t}M_0, (1-\bar{\alpha}_t)\mathbf{I}). \quad (3)$$

For training the denoising network ϵ_θ, the model predicts noise $\epsilon \sim N(0, \mathbf{I})$ by using the input M_t. The target function during the training process is as follows:

$$\mathcal{L} = \mathbb{E}_{M,\epsilon\sim\mathcal{N}(0,\mathbf{I})}\|\epsilon - \epsilon_\theta(M_t, t)\|_2^2. \quad (4)$$

U-Net [18] is typically chosen as the denoising network ϵ_θ, and during the denoising process, the network often removes many of the personal styles. Thus, we propose a method that can retain the detailed vital information as well as eliminate the white noise in the synthesized speech. More precisely, the Style-Enhanced Normalization (STEN) uses two types of information, i.e., from the text encoder (TE) and the personal styles (PS). Given the hidden vector h - a transformation of the input mel-spectrogram, we apply the layer normalization first:

$$y = \frac{h - \mathbb{E}[h]}{\sqrt{Var[h]}} * \gamma + \beta, \quad (5)$$

where γ, β are learnable parameters during the training process. Next, we augment the text representation information:

$$y = y + \text{Conv}(TE). \quad (6)$$

Then we use the personal style (PS) vector to enhance the adaptability of the model:

$$y = f_g(PS) * y + f_h(PS). \quad (7)$$

Here, f_g and f_h are the fully connected layer that performs scaling and shifting operations on the hidden vector y consecutively. By applying the STEN in each residual block of U-Net, the denoising module ϵ_θ can synthesize output that is more similar to the reference audio in only a few seconds.

In the inference procedure, the model takes the TTS output as an input $\widehat{M}_{out}$ and tries to denoise it by K time steps as shown in Figure 2-b.

Table 1: *MOS and SMOS for SEEN and UNSEEN speakers in 3 seconds.*

		VN		JA		ZH		ID		EN	
		MOS	SMOS	MOS	SMOS	MOS	SMOS	MOS	SMOS	MOS	SMOS
SEEN Speaker	Ground truth	4.65	4.72	4.8	4.66	4.95	4.81	4.69	4.60	4.37	4.58
	StyleSpeech	2.78	3.14	2.76	3.34	**3.19**	3.25	2.46	3.13	**3.20**	3.34
	DiffSinger	2.79	3.30	2.64	3.48	3.13	3.41	2.65	3.36	2.11	2.54
	STEN-TTS	**2.89**	**3.34**	**2.85**	**3.68**	3.03	**3.55**	**2.78**	**3.58**	3.20	**3.79**
UNSEEN Speaker	Ground truth (VCTK)	4.40	4.52	4.40	4.52	4.40	4.52	4.40	4.52	4.40	4.52
	StyleSpeech	1.98	2.24	**2.47**	2.93	3.07	3.03	1.85	2.44	2.10	2.55
	DiffSinger	2.00	**2.39**	2.37	3.17	**3.21**	3.22	**2.21**	**2.71**	2.53	2.5
	STEN-TTS	**2.00**	2.37	2.34	**3.19**	2.73	**3.3**	2.12	2.70	**3.01**	**3.72**

3. Experiments

3.1. Dataset

Table 2: *Number of speakers and training hours in the dataset*

Language	EN	ZH	ID	JA	VN	Total
Speakers	1151	218	400	100	54	1923
Hours	245	85	44	26	26	426

We trained the STEN-TTS in five languages: English, Chinese, Indonesian, Japanese, and Vietnamese as shown in Table 2. For English, LibrisTTS-clean-100 and LibrisTTS-clean-360 [19] were selected for our experiment. We used AISHELL-3 [20] and INDspeech NEWS LVCSR [21] datasets for Chinese and Indonesian, respectively. Japanese versatile speech (JVS) [22] is used since it is a high-quality corpus for Japanese. Finally, Vietnamese is a dataset comprising 54 speakers in both Northern and Southern Vietnam. After combining these datasets, we split them into two parts, training and validation. Validation has all languages, like the training set, and is used as seen speakers to check the performance of our proposed method. Next, CSTR voice cloning toolkit (VCTK) [23] was chosen to check the adaptation of the system for unseen speakers, since the VCTK Corpus includes 110 English speakers with different accents.

3.2. Implementation Settings

First, all audio in all languages was re-sampled to 22 050 Hz by the Librosa library[1]. Next, an espeak library[2] was used to convert text to phonemes-level. Then, all audio and their corresponding phoneme transcripts were passed to an MFA tool[3] for time-aligning. To convert the audio signal to mel-spectrogram, we use a filter length of 1024, a hop length of 256, and a window size of 1024. All of our experiments were trained in 90,000 iterations on two NVIDIA A100 GPUs. The batch size was configured to 64, we used Adam optimizer [24] in beta versions (0.9, 0.98), and the initial learning rate was 0.001. The total trainable parameter of STEN-TTS is 51M. In addition, we reproduced StyleSpeech [3] and DiffSinger [15] based on the configurations reported by the authors in their papers and added some specific modules to deal with multilingual adaptive training.

[1]https://github.com/librosa/librosa

[2]https://github.com/espeak-ng/espeak-ng

[3]https://github.com/MontrealCorpusTools/mfa-models

3.3. Evaluation Settings

We collected speakers and sentences from the validation set. More specifically, three speakers from each of the five languages were chosen randomly, and we got 15 speakers in total. For the text, 20 English sentences were selected and translated into others. Finally, speeches were synthesized for all possible combinations of speakers and sentences above, described in Figure 1. Note that we only capture 3 seconds from the audio for this evaluation.

For subjective evaluation, 10 native speakers of each language participated, including Vietnamese, Chinese, Japanese, and Indonesian, for a total of 40 participants. Each participant only listened to synthesized speech based on his/her mother language. In addition, we divided the number of synthesized English audio equally among the 40 people. After listening to the synthesized speech, the participants rated it from 1 to 5 for MOS (mean opinion score), which indicated very bad and very natural output, respectively. Next, they listened to the audio reference to check how similar the input and synthesized voices were on the SMOS (similarity MOS) scale, with 1 as very different and 5 as identical. Furthermore, we used speaker verification and speaker visualization as an objective evaluation to better understand our model.

4. Results and Discussion

4.1. Evaluation for seen and unseen speakers

Table 1 shows the results of three different models on two metrics, MOS and SMOS. Note that for ease of display we use the following abbreviations: Vietnamese (VN), Japanese (JA), Chinese (ZH), Indonesian (ID), and English (EN). For the column labeled Vietnamese (VN), we used many speakers from the five languages (VN, JA, ZH, ID, EN) to synthesize Vietnamese. After obtaining the results from participants, we statistically and carefully calculated the values for each language. Although our proposed method surpasses the previous research on MOS, the outcome is not significantly different for the seen speaker. Our proposed model has some common modules with StyleSpeech and Diffusion, and thus MOS is around 3 for three different models. However, by using STEN in the diffusion model, it considerably outperforms the other models in similarity results (SMOS). Most languages achieved approximately 3.5, which is higher than the other studies, indicating that STEN improves retention of personal styles during the denoising process of the diffusion model.

Table 3: *SMOS for cross-lingual output of 3 seconds*

	StyleSpeech					DiffSinger					STEN-TTS				
	VN	JA	ZH	ID	EN	VN	JA	ZH	ID	EN	VN	JA	CN	ID	EN
VN-to*	3.51	3.23	3.10	3.33	**3.66**	**3.75**	**3.73**	**3.86**	3.42	2.77	3.63	3.5	3.8	**3.48**	3.61
JA-to*	3.33	2.13	2.73	2.97	2.82	**3.36**	**2.30**	**3.00**	2.72	2.16	3.21	2.26	2.76	**2.93**	**3.55**
ZH-to-*	3.18	3.63	3.53	3.21	3.25	3.27	3.56	3.73	3.42	2.66	**3.39**	**3.93**	**3.76**	**3.78**	**3.94**
ID-to*	3.03	2.96	3.10	3.00	3.41	**3.24**	3.53	3.13	3.30	2.55	3.21	**3.56**	**3.70**	**3.48**	**3.44**
EN-to*	2.60	2.66	3.00	3.12	3.66	**3.36**	3.30	2.93	3.30	2.66	3.06	**3.33**	**3.33**	**3.39**	**4.00**

We then investigated the ability of our proposed method with unseen speakers. We randomly selected speakers from the VCTK dataset and used the same text as that used in the evaluation of seen speakers for each language. Table 1 shows that the proposed model considerably outperforms the other models in similarity result (SMOS) for most synthesized languages, namely Japanese, Indonesian, and English.

4.2. Cross-Evaluation

Next, we evaluated the ability of our approach in terms of transferring cross-lingual data from the subjective test. The first row of Table 3 indicates SMOS scores when the Vietnamese speakers are transformed into the five languages listed in the column, as done for the remaining languages. The results show that our method surpasses StyleSpeech and DiffSinger when performing a cross-lingual task. The SMOS scores reached nearly 3.5 for Chinese, Indonesian, and English.

4.3. Similarity Accuracy

Table 4: *SIM accuracy for three different durations*

	3 seconds	5 seconds	10 seconds
StyleSpeech	0.75	0.75	0.741
DiffSinger	0.759	**0.806**	0.778
STEN-TTS	**0.824**	0.796	**0.824**

We also conducted an objective evaluation with similarity metrics on the seen dataset. We used the ECAPA-TDNN [25] toolkit implemented by SpeechBrain for speaker verification. More specifically, after synthesizing data to other languages, the reference speaker and the synthesized speaker are input to the ECAPA-TDNN [25] model to evaluate the similarity. The output of ECAPA-TDNN is a binary value, and the prediction is 1 if the two speakers are evaluated as the same speaker and 0 otherwise. Three different durations, 3, 5, and 10 seconds, were used as the reference input for this evaluation. STEN-TTS has an accuracy of about 80% in three settings as shown in Table 4. StyleSpeech and DiffSinger's accuracy for 3 seconds is around 75%, but STEN-TTS still maintains an accuracy of 82.4%, notably higher than the other models.

4.4. Speaker Visualization

In order to obtain better insight into the effectiveness of STEN-TTS and other models, we analyzed the synthesized speech of 3 seconds and its ground truth. Specifically, we continued to use the ECAPA-TDNN model to extract the embedding of speakers, and the dimension here is 192. To visualize this embedding, we passed it to the t-SNE algorithm [26] to reduce the size, with the results displayed in Figure 3. Note that each symbol represents a different language and each color represents a different speaker. We can see that STEN-TTS can clearly separate speakers, and each speaker's pair of ground truth and multi-lingual synthesis remains in the same cluster, while in contrast, these pairs in StyleSpeech tend to become disperse. Moreover, in DiffSinger, all results for speakers JP, ZH, and ID appear very close to each other.

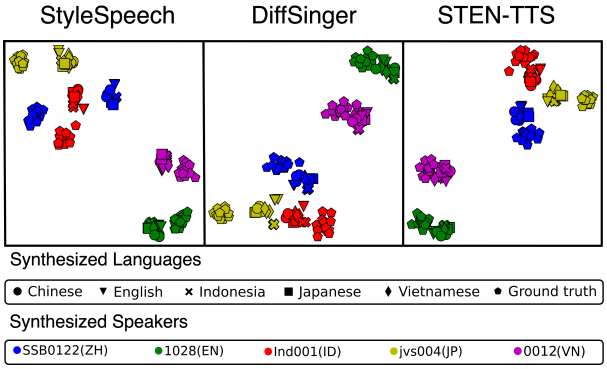

Figure 3: *Visualized embedding of synthesized and ground truth speaker in the seen dataset by t-SNE*

5. Conclusions

In this work, we demonstrated STEN-TTS, a method that replicates a speaker's style with only 3 seconds of audio reference and then synthesizes it into multiple languages. Our proposed method enhances personal styles for both seen and unseen speakers by using the Style-Enhanced Normalization module in the diffusion process. We conducted both subjective and objective evaluations, and the STEN-TTS was proven capable of achieving good results compared to previous models. Although the synthesized voice of our approach is similar to the ground truth voice, the MOS result does not show clear superiority because it has some common modules with the previous studies. Furthermore, STEN-TTS suffers from time inference because the diffusion model needs to disentangle many steps in the reverse process before returning the final result. Accordingly, we plan to analyze and solve this limitation in future work to improve the existing work.

6. Acknowledgements

Part of this work is supported by JSPS KAKENHI Grant Numbers JP21H05054 and JP21H03467.

7. References

[1] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie, C. Gulcehre, A. van den Oord, O. Vinyals, and N. de Freitas, "Sample efficient adaptive text-to-speech," in *International Conference on Learning Representations*, 2019.

[2] Y.-Y. Chou, H.-T. Lin, and T.-L. Liu, "Adaptive and generative zero-shot learning," in *International conference on learning representations*, 2021.

[3] 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*. PMLR, 2021, pp. 7748–7759.

[4] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Qin, "MultiSpeech: Multi-Speaker Text to Speech with Transformer," in *Proc. Interspeech 2020*, 2020, pp. 4024–4028.

[5] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[6] Y. Wu, X. Tan, B. Li, L. He, S. Zhao, R. Song, T. Qin, and T.-Y. Liu, "AdaSpeech 4: Adaptive Text to Speech in Zero-Shot Scenarios," in *Proc. Interspeech 2022*, 2022, pp. 2568–2572.

[7] 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 *ICASSP*. IEEE, 2018, pp. 4779–4783.

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

[9] 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 *Proc. Interspeech 2019*, 2019, pp. 2080–2084.

[10] 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. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[11] 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*. PMLR, 2021, pp. 5530–5540.

[12] J. Chevelu and D. Lolive, "Do not build Your-TTS training corpus randomly," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 350–354.

[13] H. Cho, W. Jung, J. Lee, and S. H. Woo, "SANE-TTS: Stable And Natural End-to-End Multilingual Text-to-Speech," in *Proc. Interspeech 2022*, 2022, pp. 1–5.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[15] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "DiffSinger: Singing voice synthesis via shallow diffusion mechanism," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 020–11 028.

[16] S. Liu, D. Su, and D. Yu, "DiffGan-TTS: High-fidelity and efficient text-to-speech with denoising diffusion gans," *arXiv preprint arXiv:2201.11972*, 2022.

[17] J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[19] 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 *Proc. Interspeech 2019*, 2019, pp. 1526–1530.

[20] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A Multi-Speaker Mandarin TTS Corpus," in *Proc. Interspeech 2021*, 2021, pp. 2756–2760.

[21] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, "Development of Indonesian large vocabulary continuous speech recognition system within A-STAR project," in *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*, 2008. [Online]. Available: https://aclanthology.org/I08-8004

[22] 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.

[23] C. Veaux, J. Yamagishi, and K. MacDonald, "Superseded - CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[26] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.