X-E-Speech: Joint Training Framework of Non-Autoregressive Cross-lingual Emotional Text-to-Speech and Voice Conversion

Anonymous

Abstract

Large Language Models (LLMs) have been widely used in cross-lingual and emotional speech synthesis, but they require extensive data and retain the drawbacks of previous autoregressive (AR) speech models, such as slow inference speed and lack of robustness and interpretation. In this paper, we propose a cross-lingual emotional speech generation model, X-E-Speech, which achieves the disentanglement of speaker style and cross-lingual content features by jointly training non-autoregressive (NAR) voice conversion (VC) and text-to-speech (TTS) models. For TTS, we freeze the style-related model components and fine-tune the content-related structures to enable cross-lingual emotional speech synthesis without accent. For VC, we improve the emotion similarity between the generated results and the reference speech by introducing the similarity loss between content features for VC and text for TTS.

Index Terms: joint training, text-to-speech, voice conversion, cross-lingual, emotional

1. Introduction

Text-to-speech (TTS) and voice conversion (VC) are two classic tasks in speech generation. The goal of TTS is to synthesize human-like speech based on input text[1], while VC takes speech as input to alter speech characteristics such as speaker[2] or emotion[3]. In this study, we achieve TTS and VC tasks through joint training. Due to their different nature, these two tasks use different encoders to process the input text or feature, but they share the same decoder, which is responsible for adjusting the speaker style of the generated speech.

End-to-end non-autoregressive (NAR) speech generation models, such as VITS[4] and FreeVC[5], have demonstrated the ability to produce high quality speech rapidly. In addition, joint training of TTS and VC models has been explored in studies such as HierSpeech[6], HierVST[7] and HierSpeech++[8], resulting in improved naturalness and expressiveness of synthesised speech. However, these efforts have only focused on monolingual tasks without emotion imitation. In this paper, we focus on more complex applications and aim to address two challenges simultaneously:

- Cross-lingual speech synthesis, where the language of reference speech is different from the input text (TTS) or source speech (VC).
- Emotional speech synthesis, where the generated speech needs to simulate both the speaker identity and emotional style of the reference speech.

For speech generation tasks involving complex scenarios, autoregressive (AR) approaches have achieved good results, such as VALL-E[9] and Vall-E-X[10], which can imitate speakers and emotions from reference speech in different languages to generate speech. In addition, SpeechGPT-Gen[11] and USLM[12] have also implemented multilingual TTS and VC two tasks simultaneously using Speech Large Language Model (SLLM). AR methods typically involve large-scale models trained on large amounts of data, which have excellent generalisability and perform complex tasks such as cross-lingual and emotional synthesis. However, the drawbacks of AR methods are also apparent: weaker stability and slower inference speed compared to non-autoregressive (NAR) speech synthesis methods.

There is also research into the use of NAR-based speech synthesis models for cross-lingual or emotional tasks. For example, SANE-TTS achieves cross-lingual TTS by feeding language embeddings into a text encoder[13]. PERIOD-VITS[14] and ZSE-VITS[15] imitate emotions from reference speech by receiving speaker style embeddings as input to VITS. There are also speech generation models that simultaneously handle cross-lingual and emotional tasks, such as DiCLET-TTS[16] and METTS[17] for TTS, and ConsistencyVC[18] for VC. However, there is still a lack of research exploring whether NAR-based cross-lingual emotional TTS and VC can be achieved simultaneously through joint training.

For the speech generation task, joint training is useful because the content features input to the VC model contain more information than the text input to the TTS model. As a result, the generated speech from VC models has fewer problems such as accents in cross-lingual tasks[18]. The introduction of cross-lingual content information from VC can improve the synthesis quality of cross-lingual TTS. On the other hand, joint training supports the VC task by refining cross-lingual content features to be more explicit. Through joint training with TTS, residual emotional and prosodic information in the content features can be reduced, leading to better imitation of emotional aspects from reference speech.

In this research, we introduce the joint training framework X-E-Speech. Audio samples, source code, and pretrained models are available[1]. Our contributions are summarized as follows:

- By using NAR methods, we achieve cross-lingual speaker style imitation in TTS, providing better inference speed compared to AR methods.
- By training VC and TTS together, we separate the modules responsible for speaker style and cross-lingual content in the speech synthesis model.
- We achieve cross-lingual emotional TTS and VC. Furthermore, the jointly trained VC model in conjunction with the TTS model improves the ability to imitate the emotional as-

[1]Anonymous github repository: `https://github.com/X-E-Speech/X-E-Speech-code`

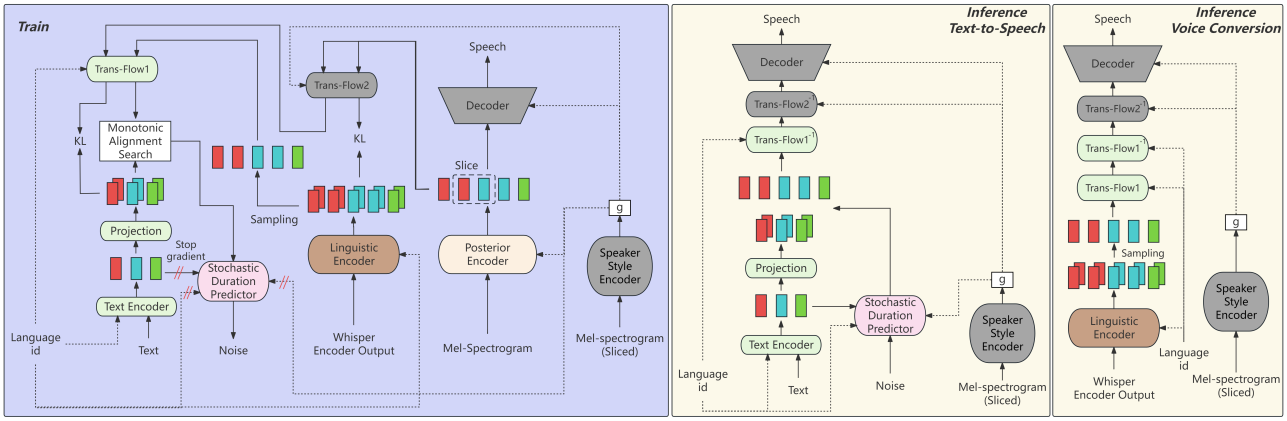

Figure 1: *System diagram showing the training and inference procedure of text-to-speech and voice conversion. Here g denotes speaker embedding.*

pect of a reference speech.

2. Method

2.1. Motivation and strategy

X-E-Speech is inspired by Hierspeech[6], ConsistencyVC[18] and VALL-E-X[10] respectively. Its backbone structure is similar to that of Hierspeech, a synthesiser that can be used for both TTS and VC tasks. We chose Hierspeech as our basic framework because it uses a hierarchical approach to achieve NAR TTS and VC simultaneously.

However, to address cross-lingual speech generation, we introduce elements of the cross-lingual voice conversion model ConsistencyVC. Specifically, we select the Whisper Encoder Output (WEO) as an intermediate feature, which is the output of the transformer encoder blocks of the Whisper model[19]. In addition, to mimic the emotions present in the reference speech, we take inspiration from VALL-E-X. We feed six-second segments of reference speech into the speaker style encoder to guide the generation of speech with the desired speaker and emotion characteristics.

Furthermore, for cross-lingual emotional speech synthesis tasks, we design three KL divergence losses to guide the training of variational inference models. To minimise accents and improve speech intelligibility, we develop a fine-tuning strategy adapted to each language. This strategy improves speech quality while preserving the similarity of the speaker in the other language.

2.2. Model architecture

As shown in Figure 1, the model architecture of X-E-Speech can be divided into two parts:

- Speaker-related: Decoder,Trans-Flow2, Speaker Style Encoder, Posterior Encoder;
- Content-related: Text Encoder, Linguistic Encoder, Stochastic Duration Predictor, Trans-Flow1.

Some of the module structure of X-E-Speech is similar to VITS and Hierspeech, although there are differences in input and model structure details. The decoder structure of the model is similar to that of VITS, which is the generator of Hifi-GAN. The difference between the text encoder and duration predictor of the model and those of VITS is the addition of a language embedding. Due to the cross-lingual task, the input to all of the content-related model structure includes language embeddings, which are obtained in a similar way to the speaker embeddings in VITS, using a linear layer.

Similarly, the style embedding obtained from the speaker style encoder is fed into all speaker-related model structures. The success of ConsistencyVC in cross-lingual VC shows that in cross-lingual speech generation, using WEO as content feature input allows the speech generation model to extract rhythmic information from WEO and voice characteristics from style embedding. The speaker style encoder uses LSTM[20] and takes randomly sliced six-second mel spectrograms of the speech as input. If the speech length is less than six seconds, it is repeated until it reaches six seconds. This further helps to prevent the leakage of content-related information from the style embedding.

The structure of the Linguistic Encoder and Posterior Encoder is similar to that of the Posterior Encoder in VITS. However, the linguistic encoder takes the WEO as input, and the input to the posterior encoder changes from spectrogram to mel-spectrogram.

Unlike VITS, X-E-Speech has two flow structures, and the Trans-Flow structure is inspired by VITS2[21], which uses normalising flows with the transformer block. This improves the ability of the flow to capture long-term dependencies, which is important for the flow to perform more complex tasks. Trans-Flow1 is responsible for encoding the rhythmic patterns from the text to the cross-lingual content feature, while Trans-Flow2 handles the speaker features from the content feature to the mel-spectrogram latent variable.

2.3. Training strategy

The training strategy is similar to VITS, using variational inference and adversarial training. However, since the model structure is divided into content-related and speaker-related parts, the training is also divided into two stages: in the first stage, the model is trained on a multilingual dataset to learn the expression of speaker characteristics and emotional features from different languages. In the second stage, some of the speaker-related model structures are frozen and the remaining parts are fine-tuned using monolingual datasets from the language of the target text. In addition, the weight of KL loss varies at different stages of training.

2.3.1. The first training stage

For the generator part, the final loss is similar to VITS, can be expressed as:

$$\mathcal{L}_{vae} = \mathcal{L}_{recon} + \mathcal{L}_{kl} + \mathcal{L}_{dur} + \mathcal{L}_{adv}(G) + \mathcal{L}_{fm}(G). \quad (1)$$

However, as shown in the first figure in Figure 1, there are three KL divergence losses among the latent variable distributions from text, WEO, and mel-spectrogram, which are mapped to each other by two Trans-flow structures.

$$\mathcal{L}_{kl} = \alpha\mathcal{L}_{kl1} + \beta\mathcal{L}_{kl2} + \gamma\mathcal{L}_{kl3}, \quad (2)$$

$$L_{kl1} = \log q_\phi(z|x_{mel}) - \log p_\theta(z|c_{text}, A), \quad (3)$$

$$L_{kl2} = \log q_\phi(z|x_{mel}) - \log p_\theta(z|x_{weo}), \quad (4)$$

$$L_{kl3} = \log q_\phi(z|x_{weo}) - \log p_\theta(z|c_{text}, A). \quad (5)$$

The definition of the prior distribution $\log p_\theta$ and the posterior distribution $\log q_\phi$ is similar to that in VITS and FreeVC. The x_{mel} is the mel-spectrogram input to the posterior encoder. The x_{weo} is the WEO input to the linguistic encoder. The c_{text} is the phonemes input to the text encoder and the A is the alignment between phonemes and latent variables from the WEO. During the forward process, Trans-Flow2 is utilized twice for calculating the L_{kl1}andL_{kl2}. Due to the L_{kl3}, which helps reduce speaker-irrelevant information in the latent variables of WEO, the model trained after the first training stage can be employed for cross-lingual emotional voice conversion tasks. The inference pipelines are shown in the third figure in Figure 1, Trans-Flow1 is utilized twice to extract the content feature.

2.3.2. The second training stage

In the second training stage, the Decoder, Trans-Flow2 and Speaker Style Encoder are frozen and the other parts are fine-tuned using a monolingual dataset. This is done to allow the Text Encoder and Stochastic Duration Predictor, which are trained on multilingual data, to focus more on the single language in the second stage. Furthermore, by freezing the speaker-related part, the fine-tuning in the second stage will not affect speaker similarity in cross-lingual speech synthesis. The posterior encoder is not frozen because it is not used in the inference period.

The α, β, γ in the $\mathcal{L}_{kl}$ are changed in the second training stage, which we will talk about in section 3.1. The model trained after the second training stage can be used for cross-lingual TTS and cross-lingual emotional TTS.

3. Experiment

3.1. Experimental setups

A sampling rate of 16,000 Hz is used for our experiments. The utterances of each speaker are randomly divided into training and test sets in a ratio of 9:1. 80-band mel-spectrograms are computed using short-time Fourier transform, with FFT, window, and hop size set to 1280, 1280, and 320, respectively. The upsampling scales of the four residual blocks in the decoder are [10, 8, 2, 2]. To avoid potential checkerboard artifacts[22], kernel sizes of [20, 16, 4, 4] are used. Our models are trained on a single NVIDIA 3090 GPU for up to 600k steps, with a batch size of 28 and a maximum segment length of 512 frames. We employ the AdamW optimizer [23] with the same weight decay and learning rate as VITS. The WEO is sourced from the large-v2 version of Whisper.

For cross-lingual TTS, the weights α, β and γ are set to [0.1,1,0.1] in the first training stage, emphasising learning of speaker characteristics. In the second stage, they are set to [1,1,1] to improve pronunciation accuracy in the target language. For the cross-lingual emotional TTS and VC tasks, the weights are set to [1,0.45,0.4] in the first stage to facilitate the learning of emotional information from the output of the speaker style encoder. In the second stage they are set to [1,1,1]. The synthesized speech used in the subjective and objective experiments is in English, with English text input, while the reference speech is in other languages. The results of the model fine-tuned with Chinese data are available on the demo page.

3.2. Cross-lingual text-to-speech

3.2.1. Dataset

In the cross-lingual text-to-speech experiment, we used trilingual datasets, including Aishell-3[24], JVS[25], VCTK[26], containing speech samples in English, Chinese and Japanese.

For the grapheme-to-phoneme conversion, we used different methods for different languages: IPA representations were obtained from espeak[2] for English, Chinese Pinyin representations were obtained from jieba[3] and pypinyin[4] for Chinese, and Japanese pronunciations were obtained from pyopenjtalk[5] for Japanese.

3.2.2. Baselines

We select two baseline cross-lingual TTS models to compare with our proposed model.

- Yourtts: An end-to-end NAR speech synthesis model trained on English, Portuguese and French using a pre-trained speaker encoder [27].
- VALL-E-X(RE): Due to the lack of official open-source models for VALL-E-X, we chose an unofficial open-source version trained on Chinese, Japanese and English datasets[6].

This comparison is unfair because the three models were trained on different datasets and have different goals. Their primary goal is zero-shot learning, and cross-lingual synthesis is just one of their capabilities. However, due to the lack of open-source models for cross-lingual speech synthesis, we had to choose these two as baselines.

3.2.3. Subjective evaluation

In the subjective experiments, we used the Mean Opinion Score (MOS) with 95% confidence intervals as a subjective metric to evaluate the naturalness and speaker similarity of the synthesized utterances. We recruited 20 native English speakers from Amazon Mechanical Turk to rate the naturalness of the speech and 13 to rate the speaker similarity between the synthesized speech and the reference speech. The speech texts were in English, while the reference speech was either Chinese from Aishell3 or Japanese from JVS. Each subject evaluating naturalness was required to assess the naturalness of 14 ground truth and 90 synthesized utterances. Each subject evaluating similarity was tasked with evaluating the similarity of 72 synthesized utterances to the utterances of the target speakers.

[2] https://espeak.sourceforge.net/
[3] https://github.com/fxsjy/jieba
[4] https://github.com/mozillazg/python-pinyin
[5] https://github.com/r9y9/pyopenjtalk
[6] https://github.com/Plachtaa/VALL-E-X

Table 1: *Results of cross-lingual text-to-speech*

	N-MOS	S-MOS	Resemblyzer	WER	CER	RTF
X-E-Speech	4.04±0.06	3.60±0.11	72.63%	13.53%	6.28%	0.024
VALL-E-X(RE)	3.50±0.07	3.50±0.11	74.50%	22.89%	12.63%	3.18
YourTTS	3.46±0.08	2.17±0.12	58.85%	9.07%	3.54%	0.023
Ground Truth	4.09±0.09	-	-	5.77%	1.89%	-

The experimental results for naturalness in Table 1 indicate that the synthesized speech has excellent naturalness and cross-lingual speaker similarity. Notably, Yourtts exhibited poor speaker similarity, possibly due to the different languages used in its training dataset.

3.2.4. Objective evaluation

We synthesized 500 utterances for the objective experiments. For the objective experiments, we conducted tests on speaker similarity, intelligibility, and inference speed. Speaker similarity was measured using the Resemblyzer tool to score the similarity between the synthesized and reference speech[7]. Intelligibility is scored by the word error rate (WER) and character error rate (CER) [28]. WER and CER are obtained using the Whisper medium.en model.

Inference speed was measured in terms of real-time factor (RTF) on an NVIDIA GeForce RTX 3090 GPU. RTF is defined as (time taken to synthesize speech) / (duration of the synthesized speech).

The experimental results in Table 1 reveal that due to the smaller scale of the VCTK dataset used for X-E-Speech compared to the LibriTTS dataset used by Yourtts, the intelligibility of the proposed model is lower than that of Yourtts. The NAR TTS model significantly outperforms AR-based TTS models in terms of inference speed. But AR structure can imitate reference speaker well.

3.3. Cross-lingual emotional text-to-speech and voice conversion

3.3.1. Dataset and baselines

For the Cross-lingual emotional text-to-speech and voice conversion task, we conducted experiments using the Emotional Speech Dataset (ESD), which contains emotional speech data from 20 speakers in English and Chinese. Since VALL-E-X can imitate the emotion of the reference speech, it remains the baseline for the cross-lingual emotional TTS task. For the Cross-lingual emotional voice conversion task, we choose the ConsistencyVC-whisper model trained on the ESD dataset as the baseline.

3.3.2. Subjective evaluation

Table 2: *Naturalness MOS with 95% confidence intervals*

	N-MOS
X-E-Speech(tts)	3.97±0.06
VALL-E-X(RE)	2.87±0.08
X-E-Speech(vc)	3.84±0.06
ConsistencyVC	3.90±0.06
Ground Truth	4.00±0.13

[7] https://github.com/resemble-ai/Resemblyzer

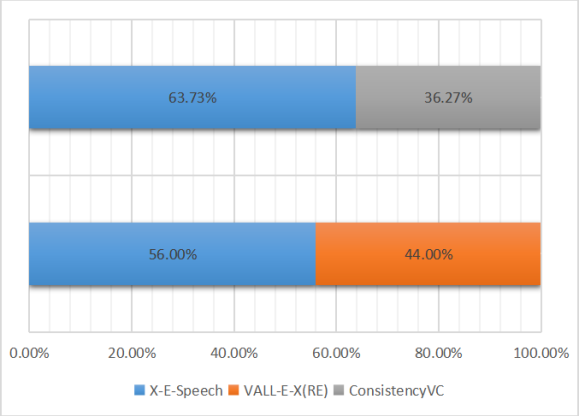

Figure 2: *The preference results for cross-lingual emotional TTS and VC.*

In the subjective evaluation, 20 native English speakers from Amazon Mechanical Turk were recruited to rate the naturalness of the synthesized emotional speech. Each subject evaluating naturalness was tasked with assessing the naturalness of 5 ground truth and 120 synthesized utterances. For the similarity task, following Du et al.[29], a preference test was conducted. Fifteen native English speakers were asked to select which speech sample was more similar to the reference speech in terms of emotion.

The results presented in Table 2 for the naturalness of TTS indicate that the synthesized emotional speech quality is better than VALL-E-X(RE). An interesting observation is that when the reference speech is highly emotional, the naturalness of VALL-E-X(RE) deteriorates significantly. The preference test results in Figure 2 demonstrate that the models can effectively mimic the emotional content of the reference speech. In terms of VC naturalness, the models perform worse than ConsistencyVC, likely due to information loss in WEO, affecting the quality of the synthesized speech. However, this loss is meaningful, as reflected in the preference test results, where the models resemble the reference speech more in terms of emotion compared to ConsistencyVC.

4. Conclusion

In this study, we proposed X-E-Speech, a joint training framework of non-autoregressive cross-lingual emotional TTS and VC. The X-E-Speech model offers faster speech generation compared to AR models due to its NAR architecture. However, NAR models face limitations, particularly in zero-shot scenarios, where their ability to mimic unseen reference speech is significantly lower compared to AR models trained on large datasets. Inspired by the achievements of BigVGAN[30], we plan to further explore the use of NAR structures to achieve zero-shot cross-lingual emotional speech generation.

5. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Interspeech 2017*, 2017.

[2] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[3] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[4] 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.

[5] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] S.-H. Lee, S.-B. Kim, J.-H. Lee, E. Song, M.-J. Hwang, and S.-W. Lee, "Hierspeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 624–16 636, 2022.

[7] S.-H. Lee, H.-Y. Choi, H.-S. Oh, and S.-W. Lee, "Hiervst: Hierarchical adaptive zero-shot voice style transfer," *arXiv preprint arXiv:2307.16171*, 2023.

[8] S.-H. Lee, H.-Y. Choi, S.-B. Kim, and S.-W. Lee, "Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis," *arXiv preprint arXiv:2311.12454*, 2023.

[9] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[10] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.

[11] D. Zhang, X. Zhang, J. Zhan, S. Li, Y. Zhou, and X. Qiu, "Speechgpt-gen: Scaling chain-of-information speech generation," *arXiv preprint arXiv:2401.13527*, 2024.

[12] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech large language models," *arXiv preprint arXiv:2308.16692*, 2023.

[13] H. Cho, W. Jung, J. Lee, and S. H. Woo, "Sane-tts: stable and natural end-to-end multilingual text-to-speech," *arXiv preprint arXiv:2206.12132*, 2022.

[14] Y. Shirahata, R. Yamamoto, E. Song, R. Terashima, J.-M. Kim, and K. Tachibana, "Period vits: Variational inference with explicit pitch modeling for end-to-end emotional speech synthesis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] J. Li and L. Zhang, "Zse-vits: A zero-shot expressive voice cloning method based on vits," *Electronics*, vol. 12, no. 4, p. 820, 2023.

[16] T. Li, C. Hu, J. Cong, X. Zhu, J. Li, Q. Tian, Y. Wang, and L. Xie, "Diclet-tts: Diffusion model based cross-lingual emotion transfer for text-to-speech—a study between english and mandarin," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[17] X. Zhu, Y. Lei, T. Li, Y. Zhang, H. Zhou, H. Lu, and L. Xie, "Metts: Multilingual emotional text-to-speech by cross-speaker and cross-lingual emotion transfer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[18] H. Guo, C. Liu, C. T. Ishi, and H. Ishiguro, "Using joint training speaker encoder with consistency loss to achieve cross-lingual voice conversion and expressive voice conversion," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[20] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[21] J. Kong, J. Park, B. Kim, J. Kim, D. Kong, and S. Kim, "Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design," *arXiv preprint arXiv:2307.16430*, 2023.

[22] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[24] 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*, 2020.

[25] 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.

[26] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.

[27] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[28] Z. Zhao, J. Liang, Z. Zheng, L. Yan, Z. Yang, W. Ding, and D. Huang, "Improving model stability and training efficiency in fast, high quality expressive voice conversion system," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 75–79.

[29] Z. Du, B. Sisman, K. Zhou, and H. Li, "Disentanglement of emotional style and speaker identity for expressive voice conversion," *arXiv preprint arXiv:2110.10326*, 2021.

[30] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022.