Cross-lingual Prosody Transfer for Expressive Machine Dubbing

Jakub Swiatkowski[*], *Duo Wang*[*], *Mikolaj Babianski*[*], *Patrick Lumban Tobing, Ravichander Vipperla, Vincent Pollet*

Amazon Science

{jswiat|duowangd|babiansk|patrilum|ravivip|vinpolle}@amazon.com

Abstract

Prosody transfer is well-studied in the context of expressive speech synthesis. Cross-lingual prosody transfer, however, is challenging and has been under-explored to date. In this paper, we present a novel solution to learn prosody representations that are transferable across languages and speakers for machine dubbing of expressive multimedia contents. Multimedia contents often contain field recordings. To enable prosody transfer from noisy audios, we introduce a novel noise modelling module that disentangles noise conditioning from prosody conditioning, and thereby gains independent control of noise levels in the synthesised speech. We augment noisy training data with clean data to improve the ability of the model to map the denoised reference audio to clean speech. Our proposed system can generate speech with context-matching prosody and closes the gap between a strong baseline and human expressive dialogs by 11.2%

Index Terms: speech synthesis, text-to-speech, prosody transfer, cross-lingual, noise-robust, automatic dubbing

1. Introduction

Intonation, stress, rhythm and style are factors of speech that are collectively referred to as prosody. To study and apply these factors for the purpose of speech generation, various acoustically inspired labelling schemes have been designed. In [1], the transplantation of prosody from an original speech clip via a system called PROTRAN was proposed. This technique involves an encoding of stylized pitch-contours and phoneme durations into a low bit-rate *enriched phonetic transcription* that can be used in conjunction with desired text to reproduce the prosody of an original recording. In our work, we circumvent the labour intensive schematizing and labeling of prosody. We adopt the term *prosody* as a general term that constitutes learned latent representations from ground truth speech audios. Similar to the definition in [2], prosody in this work refers to the encoding of variations in speech signal that remains after accounting for the variations due to phonetics, language, speaker identity, and channel effects (i.e. the recording environment and ambient noise).

In this work, we focus on cross-lingual speech synthesis for machine dubbing where the content in a source language is translated and converted into speech in a target language. Existing speech synthesis methods in machine dubbing [3, 4, 5] generate speech only based on translated text, but do not model nor transfer the expression of corresponding speech in the original language. For machine dubbing of expressive multimedia contents such as videos from various sources, it is important to convey the same emotion and expression as in the original speech [6]. In this paper, we explore cross-lingual prosody transfer for expressive speech synthesis of multimedia contents. We define cross-lingual prosody transfer as the transfer of prosody representations from speech in a source language from a source speaker to generate speech in a target language with voice characteristics of a target speaker. While exact prosody delivery varies across languages, the prosody of speech expressing the same emotions in related languages exhibits highly correlated prosody, as discussed in Section 4.6 in [6]. In our work, we explore these cross-lingual correlations for the purpose of prosody transfer. We study European languages such as English, German, French, Italian and Spanish, and focus on English-Spanish prosody transfer. In this work, we do not focus on more distant languages such as Japanese.

Cross-lingual prosody transfer brings additional context from speech in source language, but it also involves a number of challenges that are not present in conventional Text-To-Speech (TTS) solutions. First, currently cross-lingual prosody transfer has to be learned without access to multilingual parallel speech datasets due to the scarcity of such datasets. The available parallel datasets [7] lack expressivity. The absence of expressive parallel datasets also means speech-to-speech translation methods [8] are not applicable. Second, available non-parallel speech datasets lack the full range of expressivity present in human speech [9]. Therefore, we resort to gathering expressive speech of different languages and speakers from in-house multimedia data. However, such data was not recorded for the purpose of TTS systems. For example, the data contains channel noise, which needs to be alleviated to generate clean and expressive speech.

In this work, we introduce a solution based on conditional variational autoencoder with adversarial learning for end-to-end text-to-speech (VITS) [10]. Our solution learns cross-lingual prosody transfer from non-parallel data. We use parallel translated texts during inference, but our proposed system doesn't require parallel text nor parallel audio data during training. It enables the cross-lingual prosody transfer by learning prosody representations that are agnostic to speakers and languages. The prosody representations are learnt via a variational reference encoder [11] with carefully balanced regularization. The learnt representations can be transplanted from a reference audio in source language spoken by a source speaker to generate speech in target language with the voice characteristics of a target speaker. Furthermore, to improve the robustness of our model to noisy reference audio, we propose two different approaches. The first approach utilizes a noise modelling extension to our reference encoder module that disentangles the prosody and the channel noise, where a denoised signal extracted from a reference audio is utilized. On the other hand, in the second approach, we augment the training data with clean

[*] Equal contribution

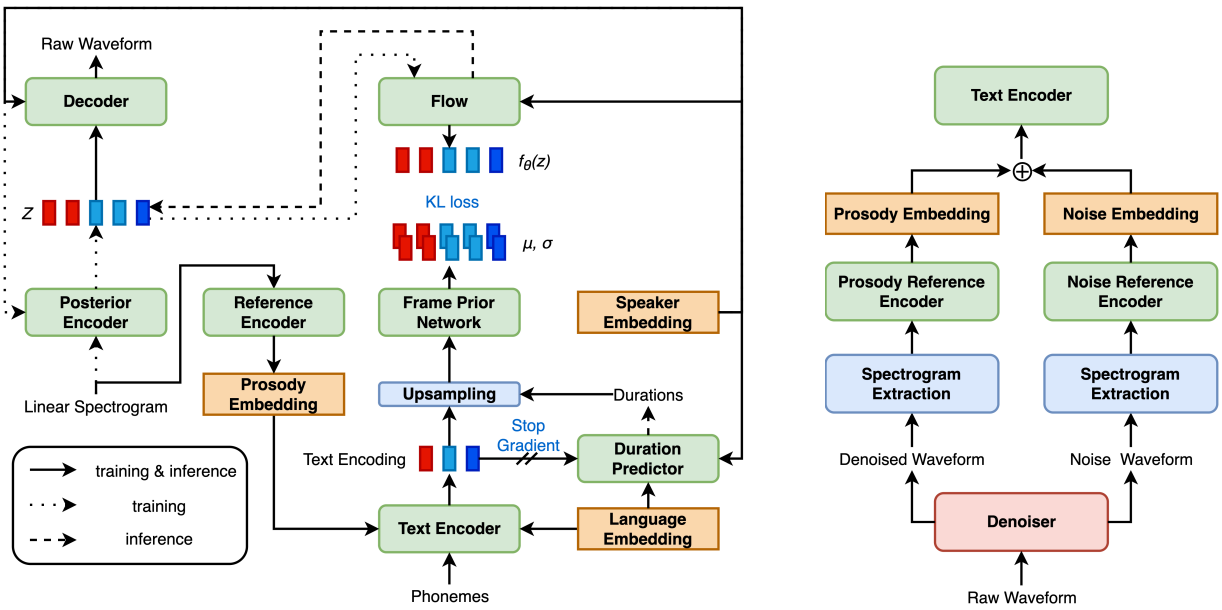

Figure 1: *Architecture diagram of the main system with prosody reference encoder (left) and explicit noise modelling method utilizing both prosody and noise reference encoders. (right).*

speech data to improve the capability of our model to map a denoised reference audio to clean speech. Both approaches allow our system to learn from noisy data and to generate high-quality clean speech in a target language even when it is provided with a noisy reference audio from the source language.

Related to our system, numerous works on the prosody transfer within a single language have been proposed, such as with the use of reference encoder [12], with style tokens [13], or with variational autoencoder (VAE) [11]. Concurrent to our work, [14] also extended VITS with a reference encoder. Cross-lingual setting was explored in [15], however, this work is focused on style transfer based on categorical labels, which are provided as ground truth during training. Last but not least, explicit noise modelling in TTS systems has also been studied [16, 17], but transferring prosody from noisy reference recordings was not explored in these studies. To the best of our knowledge, our work is the first to address the problem of cross-lingual prosody transfer for machine dubbing and is robust to noisy data. To sum up, our contributions are:

- We show that cross-lingual prosody transfer can be achieved with a multilingual model trained without parallel data.
- We propose a reference encoder architecture that disentangles prosody and channel noise allowing for clean speech synthesis from a noisy reference audio. We also investigate the augmentation of noisy data with clean training data to improve the capability of the model to map a denoised reference input to clean speech.

2. Modelling

Our proposed model consists of a backbone adapted from VITS [10], a prosody reference encoder to encode prosody information from the reference audio, and an optional noise reference encoder to model noise information. Figure 1 (left) shows an overview of our proposed model architecture.

We derive our base model by adapting the following changes from the literature. First, we replace VITS's monotonic alignment search algorithm with explicit duration predictor and extend prior encoder module with a frame prior network as in [18]. Second, we incorporate speaker embeddings and language embeddings as in [19] for training on multi-speaker and multilingual datasets. This is also depicted in Figure 1. Finally, we replace HiFiGAN decoder [20] with a BigVGAN-base decoder [21] as BigVGAN shows improved generalization performance compared to HiFiGAN. We find that these changes significantly improve over the original VITS and keep them fixed in all our experiments.

2.1. Prosody Encoder

Prosody reference encoder extracts prosody embedding from a reference speech input. As we explicitly condition speaker and language variations via respective embeddings, the prosody encoder captures the remaining variability related to prosody. The prosody embedding is used to condition the model to synthesise speech with similar prosody to the reference speech sample. Formally, prosody reference encoder can be represented as a function h that encodes speech representation s into prosody embedding e as $e = h(s)$. We can either use the output of posterior encoder or the extracted linear spectrogram as speech representation s. In our experiments, we find both to have similar performances.

In practice, the reference encoder h is parameterized by a variational encoder module that consists of five convolutional layers of 512 channel size and a kernel size of 3, and one bi-directional LSTM layer of channel size 512. The cell states of LSTM layer is then further processed by two fully connected layers that output the parameterized diagonal Gaussian distribution, that is regularized using KL-Divergence with a standard Gaussian $\mathcal{N}(0, I)$. The variational Bayesian formulation has two advantages. First, this formulation allows interpolable embedding space, which is conducive for the sampling of prosody. Secondly, carefully tuned KLD regularizes the prosody embedding to reduce speaker and language information contained within the embedding, which is essential for cross-speaker and cross-lingual prosody transfer.

We experimented with various ways of conditioning the

model on extracted prosody embedding and found that conditioning in the text encoder module (Figure 1) produces the best result. Intuitively, conditioning in the text encoder allows a joint modelling $P(c, e)$ of the text embedding c and the prosody embedding e, which makes it possible to model long-term dependencies between text sequence and prosody embedding. We denote this system as Variational Inference for Prosody Transfer (VIPT).

2.2. Noise Modelling

The prosody encoder, designed to encode the prosody of the reference audio, also encodes other artifacts such as background noise and distant speech not annotated in the text. In our empirical study, we found that the presence of these artifacts severely degrades the quality of speech synthesis. We propose two approaches to tackle this issue.

In the first approach, we introduce an explicit noise modelling method (Figure 1, right) to our system, which enables to disentangle the prosodic information from the noise information. As a result, clean speech audio can be generated even when a noisy reference audio is provided. At training time, we use an external denoiser [22] to split reference audio into denoised waveform and noise residual waveform. We feed the spectrograms extracted from the two audio streams into separate reference encoders resulting in two disentangled embeddings: a prosody embedding from denoised audio and a noise embedding from the noise residual. Finally, we concatenate the prosody and noise embeddings to condition the text encoder as described in Section 2.1. In this way, a mapping from the noise embedding to noise artifacts contained in the target waveform is learnt. At inference time, the prosody embedding is extracted from a given denoised reference audio containing the desired prosody, while the noise embedding is derived from a separate clean utterance. We denote this system as Variational Inference for Prosody Transfer with Noise Modelling (VIPT-NM).

In the second approach, we use the base VIPT architecture, but input the denoised reference audio during inference time. This approach is able to reduce the noise level in the synthesised speech, but may also introduce distortions due to unseen denoised reference audio input that are out of training data distribution. To remedy this, we add clean data as a proxy for denoised audio to our dataset and train our model with both noisy data and clean data. In this way, the out-of-distribution condition of the denoised reference audio is alleviated, which, in turns, improves the synthesis quality.

2.3. Training Setup

We follow the training setup of original VITS [10], where we include the use of short-term Fourier transform (STFT) discriminator as in BigVGAN [21]. The final loss can be expressed as follows:

$$\mathcal{L} = \mathcal{L}_{VITS} + \alpha_1 \mathcal{L}_{ProsodyKLD} + \alpha_2 \mathcal{L}_{NoiseKLD} \quad (1)$$

where $\mathcal{L}_{VITS}$ represents the VITS loss terms except with the adversarial loss changed to the BigVGAN's formulation. $\mathcal{L}_{ProsodyKLD}$ and $\mathcal{L}_{NoiseKLD}$ are KLD losses for prosody and noise reference encoders respectively. After hyper-parameter search of 12 runs, we found the best KL-Divergence loss coefficients α_1 and α_2 to be both 0.001. We use mixed precision training on eight NVIDIA V100 GPUs. The batch size is set to 30 per GPU and the models are trained up to 700k steps. The generative part of the VIPT-transfer model has a total of 90 million parameters and discriminators have 47 million parameters.

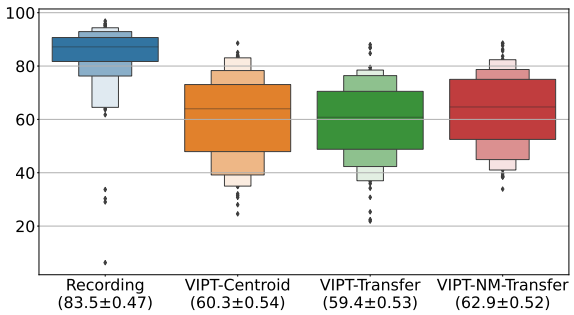

Figure 2: *Subjective listeners ratings from the machine dubbing MUSHRA test for VIPT and VIPT-NM. Values under labels represent mean scores and their respective standard errors.*

3. Evaluations

We used an internal multi-speaker multilingual dataset mined from existing in-house multimedia source data that contains expressive speech recorded in varying acoustic conditions. The dataset comprises 118 hours of speech recordings from 127 speakers in five different locales; namely US English, Castilian Spanish, French, German and Italian. Speaker age groups range from children to elderly. We split data into training, development, and test sets using a 85:5:10 ratio. For the evaluation of cross-lingual prosody transfer, we ran MUSHRA tests on a held-out subset of 100 US English utterances with expressive human dubbing in Castilian Spanish. For the subjective evaluation, for the sake of brevity, we focus on prosody transfer from US English to Castilian Spanish as a representative of other language combinations. Our proposed method also works for other language pairs, and we briefly discuss objective metrics evaluation for the other language pairs in Section 3.3. In order to provide testers with a precise context for prosody assessment, we presented the audio samples overlaid on the corresponding videos. We evaluated four systems:

- *VIPT-Centroid* - We aim to pick a baseline model that generates high quality speech, but does not have prosody transfer capability. We introduce VIPT-Centroid, which is the same as VIPT model, but uses the centroid of prosody embeddings calculated across denoised reference samples for the target speaker. VIPT-Centroid is a stronger baseline than other VITS-based external models such as YourTTS [23], because those models, without a reference encoder, tend to internalize the noise into the model parameters and frequently generate noisy speech with higher rate of mispronunciations. To illustrate this, we measured Signal-to-Noise Ratio (SNR) of VIPT-Centroid and YourTTS outputs by using the same denoiser as used in Section 2.2. VIPT-Centroid has a SNR of 45.2 dB, which is significantly higher than YourTTS's SNR ratio of 34.8 dB.
- *VIPT-Transfer* - As above but with the prosody embedding extracted from the denoised audio of a source English speaker.
- *VIPT-NM-Transfer* - As above but with explicit noise modelling applied.
- *Recording* - Professional human Spanish dubbing.

In the MUSHRA test, 25 native Spanish speakers were presented with the video samples in a random order side-by-side, and were asked to *"Rate the vocal performance in the Spanish video dubbing samples with respect to the English reference video"*. Each test case was scored by all 25 testers indepen-

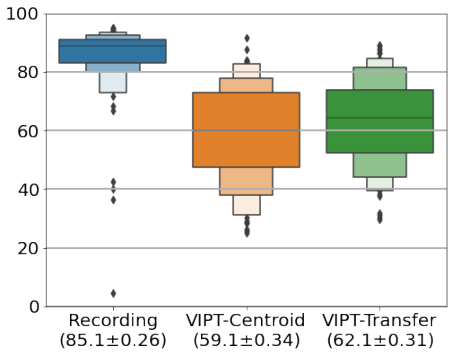

Figure 3: *Subjective listeners ratings from the cross-lingual prosody transfer MUSHRA test for VIPT-Transfer with additional clean training data. Values under labels represent mean scores and their respective standard errors.*

dently.

3.1. Perceptual Metrics

Figure 2 shows that while VIPT-Transfer on average achieved lower MUSHRA scores than the baseline VIPT-Centroid, VIPT-NM-Transfer was significantly better. On closer inspection, we observed that the VIPT-Transfer system scored lowest for utterances with particularly noisy English reference audios, thereby dragging down the mean score despite of its capability to perform prosody transfer. VIPT-NM-Transfer was more robust to the negative impact of the noise in reference audio and resulted in better matching prosody than the baseline VIPT-Centroid, thereby achieved a statistically significant MUSHRA score increase and closed the gap to human dubbing by 11.2%.

Additionally, we evaluated the effects of adding clean data to improve the synthesis quality when using denoised reference audio. For model training, we added 480 hours of internal clean speech data that consists of 183 additional speakers within the same 5 locales as the original data. Similarly as before, on the cross-lingual prosody transfer from English to Spanish, we used denoised English reference audio to condition the prosody encoder for synthesising Spanish speech with the desired prosody. For the perceptual metric evaluation, we conducted a MUSHRA test with the same setup as above. Figure 3 shows the MUSHRA scores of VIPT-Transfer compared to VIPT-Centroid. The improved score of VIPT-Transfer shows that adding clean data allowed us to effectively use a denoised reference audio for performing cross-lingual prosody transfer without a significant compromise in terms of the stability.

3.2. Analysis of Prosody Embedding Space

Cross-lingual prosody transfer should only transfer the prosody but not language specific accents to the target language. This requires the prosody embedding space to be disentangled from language categories. In order to verify this and to further understand the learnt VAE reference encoder embedding space, we used t-SNE to reduce dimensions of the embedding space to $\mathcal{R}^2$ and plotted randomly sampled embedding using a colored scatter plot. The embedding was taken from the output mean predicted from the reference encoder for the corresponding reference sample.

Figure 4 depicts t-SNE plots of randomly sampled utterances' prosody embedding from five different locales in our dataset with 600 utterances per locale. It can be observed that there is no significant locale clustering, which indicates that the

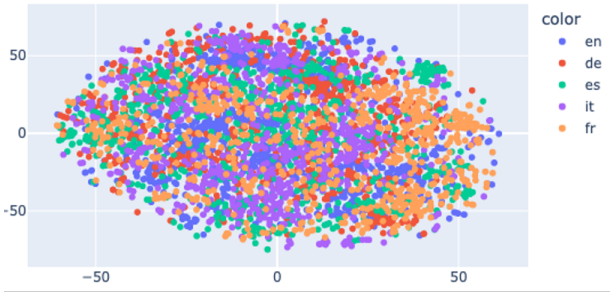

Figure 4: *T-SNE plot of VAE reference encoder embedding space coloured by languages.*

learnt reference embedding space was locale/language independent. This local-independent prosody distribution is essential for performing cross-lingual prosody transfer.

3.3. Objective Metrics For Other Language Pairs

In this section, we discuss objective metrics evaluation for language pairs other than US English and Castilian Spanish. As metric of measure prosody transfer, we focus on F0 statistics including Mean Squared Error (MSE) and Pearson correlation between synthesised speech and the corresponding Spanish human dubs. In Table 1 F0 objective metrics are computed for an external baseline YourTTS [23], and VIPT-Transfer different language pairs on the same test set used for the MUSHRA evaluation. It can be seen that VIPT-Transfer for any included language pair outperforms YourTTS in terms of F0 metrics, which indicates that our proposed method works for more than one language pair. The VIPT-Transfer-En-To-Es system gives the best scores for both metrics, which we hypothesise is due to higher proportion of English utterances in our training data.

Table 1: *F0 Metrics comparing an external baseline YourTTS [23] and our VIPT-Transfer model with prosody transfer for different language pairs. Mean Squared Error (MSE) and correlation coefficient are computed against corresponding human Spanish recordings.*

System	MSE ↓	Correlation ↑
YourTTS	8367.7	0.30
VIPT-Transfer-En-To-Es	6970.0	0.40
VIPT-Transfer-It-To-Es	7639.2	0.35
VIPT-Transfer-Fr-To-Es	7724.7	0.33
VIPT-Transfer-De-To-Es	7693.4	0.34

4. Conclusions

We presented a novel solution that learns cross-lingual prosody transfer from non-parallel noisy speech data. We showed that our proposed solution can generate dubbed speech with context-matching prosody. We further demonstrated two approaches to address challenges posed by noise in multimedia data. First, we introduced a novel noise modelling module that disentangles noise from prosody, where denoised signal extracted from reference audio is utilized. Second, we augment noisy data with clean training data to improve the capability of the model to map denoised reference audio to clean speech. Through subjective and objective evaluations we showed that our system outperforms a strong baseline in the task of speech generation for automatic dubbing.

5. References

[1] B. Van Coile, L. Van Tichelen, A. Vorstermans, J. Jang, and M. Staessen, "Protran: a prosody transplantation tool for text-to-speech applications," in *International Conference on Spoken Language Processing*, 1994.

[2] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, and R. A. S. Weiss, Ron J. Rob Clark, "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," in *International Conference on Machine Learning*. PMLR, 2018.

[3] M. Federico, R. Enyedi, R. Barra-Chicote, R. Giri, U. Isik, A. Krishnaswamy, and H. Sawaf, "From speech-to-speech translation to automatic dubbing," in *IWSLT 2020*, 2020.

[4] J. Matoušek and J. Vít, "Improving automatic dubbing with subtitle timing optimisation using video cut detection," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 2385–2388.

[5] J. Effendi, Y. Virkar, R. Barra-Chicote, and M. Federico, "Duration modeling of neural tts for automatic dubbing," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8037–8041.

[6] W. Brannon, Y. Virkar, and B. Thompson, "Dubbing in practice: A large scale study of human localization with insights for automatic dubbing," *Transactions of the Association for Computational Linguistics*, 2021.

[7] Y. Jia, M. T. Ramanovich, Q. Wang, and H. Zen, "CVSS corpus and massively multilingual speech-to-speech translation," in *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2022, pp. 6691–6703.

[8] A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang *et al.*, "Direct speech-to-speech translation with discrete units," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3327–3339.

[9] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003. [Online]. Available: https://aclanthology.org/2021.acl-long.80

[10] 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.

[11] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6945–6949.

[12] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.

[13] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[14] K. Mitsui, T. Zhao, K. Sawada, Y. Hono, Y. Nankaku, and K. Tokuda, "End-to-End Text-to-Speech based on latent representation of speaking styles using spontaneous dialogue," in *Proc. Interspeech*, 2022, pp. 2328–2332.

[15] D. Rattcliffe, Y. Wang, A. Mansbridge, P. Karanasou, A. Moinet, and M. Cotescu, "Cross-lingual Style Transfer with Conditional Prior VAE and Style Loss," in *Proc. Interspeech 2022*, 2022, pp. 4586–4590.

[16] C. Zhang, Y. Ren, X. Tan, J. Liu, K. Zhang, T. Qin, S. Zhao, and T.-Y. Liu, "DenoiSpeech: Denoising text to speech with frame-level noise modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7063–7067.

[17] T. Saeki, K. Tachibana, and R. Yamamoto, "DRSpeech: Degradation-Robust Text-to-Speech Synthesis with Frame-Level and Utterance-Level Acoustic Representation Learning," in *Proc. Interspeech*, 2022, pp. 793–797. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-294

[18] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, "VISinger: Variational inference with adversarial learning for end-to-end singing voice synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7237–7241.

[19] 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, pp. 1–5.

[20] 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.

[21] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A Universal Neural Vocoder with Large-Scale Training," *International Conference on Learning Representations*, 2023.

[22] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani, and A. Krishnaswamy, "PoCoNet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," 2020.

[23] 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.