CROSS-SPEAKER STYLE TRANSFER FOR TEXT-TO-SPEECH USING DATA AUGMENTATION

Manuel Sam Ribeiro, Julian Roth, Giulia Comini, Goeric Huybrechts, Adam Gabryś, Jaime Lorenzo-Trueba

Amazon Alexa, TTS Research

{manuerib, huybrech, truebaj}@amazon.com

ABSTRACT

We address the problem of cross-speaker style transfer for text-to-speech (TTS) using data augmentation via voice conversion. We assume to have a corpus of neutral non-expressive data from a target speaker and supporting conversational expressive data from different speakers. Our goal is to build a TTS system that is expressive, while retaining the target speaker's identity. The proposed approach relies on voice conversion to first generate high-quality data from the set of supporting expressive speakers. The voice converted data is then pooled with natural data from the target speaker and used to train a single-speaker multi-style TTS system. We provide evidence that this approach is efficient, flexible, and scalable. The method is evaluated using one or more supporting speakers, as well as a variable amount of supporting data. We further provide evidence that this approach allows some controllability of speaking style, when using multiple supporting speakers. We conclude by scaling our proposed technology to a set of 14 speakers across 7 languages. Results indicate that our technology consistently improves synthetic samples in terms of style similarity, while retaining the target speaker's identity.

Index Terms— text-to-speech, speaking style transfer, cross-speaker, data augmentation

1. INTRODUCTION

Text-to-speech (TTS) technology is consistently reducing the perceived gap between synthetic and natural speech. This is being achieved with the development of novel acoustic modeling [1, 2] and waveform generation techniques [3, 4]. Leveraging these methods, researchers have been focusing on the flexibility and controllability of TTS systems, especially for expressive data [5, 6, 7]. The development of flexible systems requires the ability to control speaking style and speaker identity, among other speech attributes. Speaking style denotes the global attributes that describe the emotion, affect, and/or generic attitude conveyed through speech by a speaker in a particular domain. For example, we may define speaking styles such as read, newscaster, conversational, or emotional speech. The ability to control these attributes is essential for TTS voices that are flexible and adaptable to multiple scenarios and domains. Traditionally, extending TTS voices to new domains where speaking style is relevant involved additional data collection from the target voice. This method, however, does not scale well, as it is not always feasible to record more data for a specific voice talent. An alternative approach is to transfer speaking style from other speakers for which recorded data is already available.

Cross-speaker style transfer involves the generation of speech samples that are perceived to have the identity of a target speaker and the speaking style of a supporting speaker. To control speaking style, recent work proposed the inclusion of reference encoders [8, 9]. Together with textual input, the TTS model inputs a reference speech representation that is used to condition the generated speaking style. Frequently used methods for the reference encoder are based on Global Style Tokens (GST) [9] or Variational Auto-Encoders (VAEs) [10, 11]. Although these methods could be used for cross-speaker style transfer, they have some shortcomings. They may be dependent on similar text, where the reference waveform is similar to the textual input to be synthesized [8, 12]. Or they might not model the target speech attributes, especially in disjoint corpora, without data from the target speaker in the target speaking style [13].

For these reasons, most systems for cross-speaker style transfer aim to explicitly disentangle speaker identity and speaking style from other speech attributes. This is primarily accomplished with multi-speaker TTS models [7]. For speech generation, the model is conditioned on the target speaker identity and the desired speaking style. Recent studies proposed the usage of multiple reference encoders, with each encoder modeling a specific speech attribute. These systems can be trained using intercross or adversarial training [14, 13]. Models are typically optimized on a variety of loss functions, such as cycle consistency loss [15], adversarial consistency loss [13, 16], N-pair loss [17] or other loss functions defined over latent representations of speaker identities and/or speaking styles [18, 16]. Instead of multiple encoders, hierarchical architectures were also proposed [19]. Alternatively, control of speaking style may be left to models of *f0* and duration, learned separately or implicitly with the TTS model [7, 6, 20].

A related area of research focuses on the development of TTS voices for low-resource scenarios. Such methods aim to synthesize speech given a limited amount of training data from the target speaker, domain, or language. Recent studies proposed to augment small corpora with synthetic data, either via voice conversion [21, 22] or a large text-to-speech system [23]. These studies provided evidence that high-quality synthetic data can be used efficiently to complement natural data.

In this paper, we propose to address the problem of *cross-speaker style transfer for text-to-speech using data augmentation*. We assume to have a corpus of neutral non-expressive data from a target speaker and some supporting conversational expressive data from other speakers. Our goal is to build a TTS system that is expressive, while retaining the target speaker's identity. Our method uses voice conversion to generate high-quality data from a set of supporting expressive speakers. The synthetic expressive data is pooled with the natural non-expressive data to train a single-speaker multi-style TTS system. We show that our proposed method is flexible and capable of transferring speaking style across speakers. Additionally, we show the effectiveness of our proposed method when 1) using data from one or more supporting speakers; 2) using as little as 1 hour of supporting data; 3) controlling the speaking style via the TTS reference encoder's latent space; and 4) applied to a variety of languages and speakers.

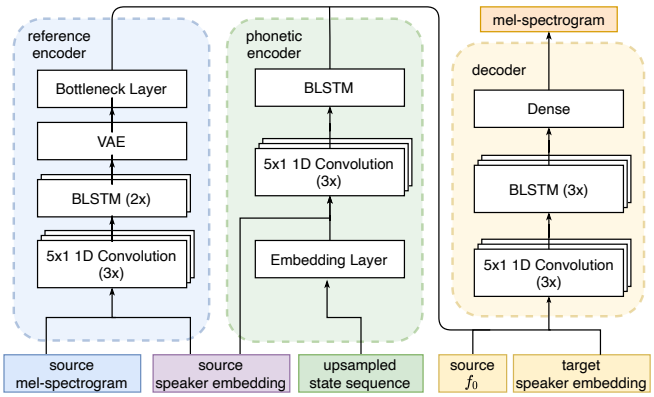

Fig. 1. Voice Conversion model architecture.

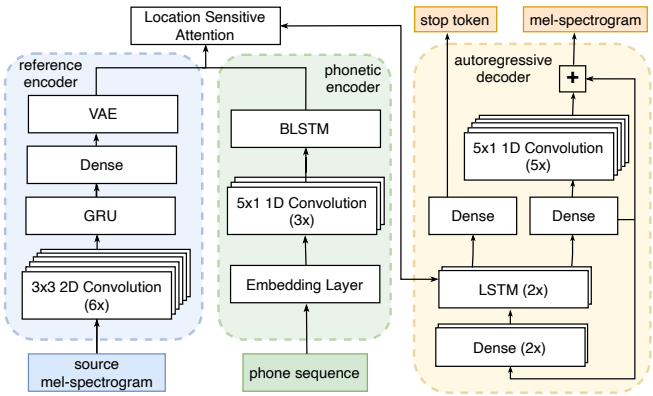

Fig. 2. Text-to-Speech model architecture.

2. MODEL ARCHITECTURE

2.1. Voice Conversion

Our voice conversion approach (Figure 1) is based on CopyCat [24], extended to be conditioned on the source utterance's *log-f0* [25]. The purpose of this architecture is to preserve linguistic content and prosodic attributes, while modifying only speaker identity. The reference encoder inputs a source speaker's mel-spectrogram and it includes a bottleneck layer that downsamples and upsamples the latent representations along the time dimension [26]. The phonetic encoder's architecture follows the Tacotron 2 text encoder [1]. It inputs a categorical representation of Hidden Markov Model (HMM) states. The HMM state sequence is found by force-aligning the training data at the phone level to the corresponding Mel-Frequency Cepstral Coefficient sequence. We use 3-state left-to-right HMMs for each individual phone. To match the time resolution of the mel-spectrogram, each time-aligned HMM state is upsampled to the frame level. Both reference and phonetic encoders further input a speaker embedding defined at the utterance level, which is broadcasted to the number of frames in each utterance [21]. Speaker embeddings are learned on a large multi-speaker multi-lingual corpus and optimized on a Generalized End-to-End Loss [27]. The encoded reference and phonetic sequences are concatenated with the utterance's *log-f0* and speaker embedding. The model is optimized on data from the target and supporting speakers for 100k steps using a batch size of 32 utterances. We use a KL-divergence loss for the VAE component and an L1 loss on the source and reconstructed mel-spectrograms. We further fine-tune the model for 25k steps only on training data from the target speaker.

For voice conversion, the reference and phonetic encoders input data from the source speaker, while the decoder is conditioned on information from the target speaker. The target speaker embedding is the centroid of all embeddings from that speaker's training data. The observed source utterance's *f0* is mean-normalized to the target speaker's mean *f0*. The last layer of the decoder transforms the input frame-by-frame, using contextual information provided by the recurrent layers. Note that there is no attention or alignment required in this architecture, as we preserve the source utterance's duration. With the addition of the source *f0* signal, the reconstructed mel-spectrogram preserves the prosodic properties describing the source style. Any other relevant information not accounted for by duration, *f0*, or speaker identity, is carried-over by the reference encoder.

2.2. Text-to-Speech

Our TTS model (Figure 2) is a sequence-to-sequence encoder-decoder architecture based on Tacotron 2 [1], using a single-head location-sensitive attention mechanism [28]. The phonetic encoder and decoder follow the architecture proposed by Shen et al [1]. The reference encoder follows the architecture proposed by Skerry-Ryan et al [8] with the addition of a Variational Auto-Encoding layer [10]. TTS systems are trained for 400k steps using a batch size of 32 utterances and optimized on an L1 loss for the generated mel-spectrograms, a KL-divergence loss for the Variational Autoencoder, and a cross-entropy loss for the stop-token. The phonetic sequence is extracted from language-specific front-ends and corresponds to phone identities, word boundary tokens, and stress markers.

For style transfer TTS systems, we pool the synthetic expressive mel-spectrograms generated by the voice conversion model with the natural non-expressive mel-spectrograms from the target speaker. Because speaker identity has been modified by voice conversion, there is no need to train multi-speaker models or use multiple reference encoders. To generate speech samples, we condition the decoder on a VAE z-vector computed over the training data converted from a single supporting speaker, corresponding to a unique speaking style. We finally convert the mel-spectrograms generated by the TTS models to time-domain waveforms using a Parallel Wavenet universal neural vocoder [29].

3. EXPERIMENTS

Our goal is to build conversational text-to-speech systems for voices that only have speech recordings read in a *neutral speaking style*. This speaking style is characterized by flat, monotonous, and unexpressive read speech. As supporting data, we use recordings from different speakers in a *conversational speaking style*, which aims to capture a natural, friendly, and expressive speaking style. Throughout our experiments, we use the terms Target, Supporting, and Source speaker. "Target speaker" denotes the speaker for which neutral data is available, and the speaker identity we wish to preserve in the synthetic samples. "Supporting speakers" indicates the set of speakers from which we draw supporting conversational data used to augment the TTS models. "Source speaker" denotes the supporting speaker used to condition the TTS model when generating speech samples. We compute a VAE z-vector centroid over the set of voice-converted training samples from the Source speaker.

System	Naturalness	Speaker Sim.	Style Sim.
Recordings	61.51 ± 1.53	-	-
Neutral	54.76 ± 1.39	71.93 ± 1.54	38.12 ± 1.65
Augmented (1 spk)	59.07 ± 1.37	64.56 ± 1.57	58.60 ± 1.58
Augmented (4 spks)	58.98 ± 1.39	65.03 ± 1.55	60.05 ± 1.55
Augmented (8 spks)	59.51 ± 1.36	64.54 ± 1.57	59.17 ± 1.57
Source Speaker	-	27.85 ± 1.61	72.81 ± 1.59

Table 1. MUSHRA evaluation in terms of Naturalness, Speaker Similarity, and Speaking Style Similarity. Results indicate mean score with 95% confidence interval. "Augmented" systems denote target speaker TTS systems augmented with conversational data from a number of supporting speakers. "Source speaker" indicates a conversational TTS system from the source supporting speaker.

3.1. Supporting speakers

To investigate our proposed cross-speaker style transfer approach, we use internal corpora of Brazilian Portuguese data, choosing the target speaker to be a female speaker and supporting speakers to be gender-balanced. *Neutral* denotes a system trained on 10 hours of neutral data from the target speaker. *Augmented* systems indicate models trained using data augmentation via voice conversion. For this experiment, we convert a total of 8 hours of conversational data to the identity of the target speaker. We keep the amount of supporting data fixed and we vary the number of supporting speakers. The system with 1 supporting speaker uses 8 hours of data from a single speaker, while the systems using 4 and 8 supporting speakers use 2 hours and 1 hour of data from each supporting speaker, respectively. The source speaker is kept constant across all augmented systems. For comparison, *Source* indicates a TTS system trained on the 8 hours of conversational data from the source speaker. We evaluate our systems with a set a MUSHRA-like (MUltiple Stimuli with Hidden Reference and Anchor) [30] listening tests. Naturalness is evaluated omitting the reference and including a recording sample with the competing systems. We further evaluate systems in terms of Speaker Similarity and Style Similarity. The speaker similarity evaluation uses a reference sample drawn from the training data of the target speaker, while the style similarity evaluation uses a reference sample drawn from the training data of the source speaker. For these tests, listeners were asked to rate the samples based only on the similarity of the speaker identity or the speaking style, ignoring all other attributes. Each listening test included 150 utterances generated by each of the competing systems. Utterances were rated by 100 native speakers using a crowdsourcing platform. Each listener rated no more than 15 MUSHRA screens.

Results in Table 1 show MUSHRA mean scores with a 95% confidence interval. We observe that the proposed Augmented systems improve naturalness over the Neutral system. In terms of speaker similarity, the Augmented systems score below the Neutral system, although this is somewhat expected due to the differences in speaking style between competing and reference samples. On average, however, the Augmented systems bridge the gap between source and target speakers by 83.6%. For style similarity, the Augmented systems outperform the neutral system, suggesting that listeners are able to discern between the speaking style generated by the competing systems. For each subjective evaluation, we perform paired two-sided t-tests on the MUSHRA scores with a Holm-Bonferroni correction for multiple comparisons. Across the three evaluated dimensions, we observe no statistically significant difference between Augmented systems at the level of $p < .01$.

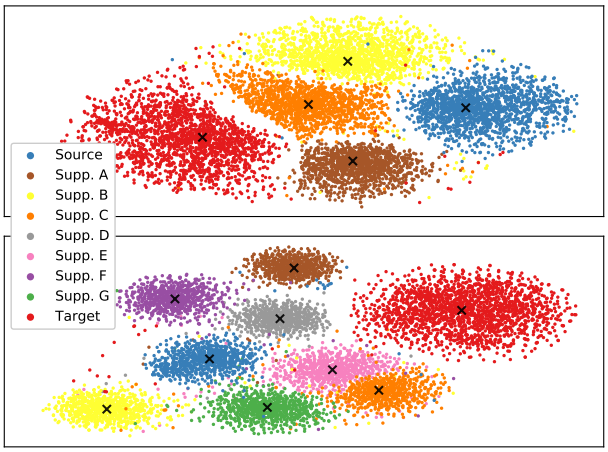

Fig. 3. VAE space for systems augmented with voice-converted data from 4 (top) and 8 supporting speakers (bottom), visualized with t-SNE [31]. The centroid of the training data converted from each speaker is marked with **X**.

	Reference				
Centroid	**Target**	**Source**	**Supp. A**	**Supp. B**	**Supp. C**
Target	**48.28%**	6.55%	14.33%	5.52%	10.74%
Source	13.45%	**32.07%**	28.33%	22.07%	25.19%
Supporting A	13.45%	8.28%	**29.34%**	7.59%	12.59%
Supporting B	10.02%	**32.41%**	14.67%	**45.16%**	17.04%
Supporting C	14.80%	20.69%	13.33%	19.66%	**34.44%**
Total	100%	100%	100%	100%	100%

Table 2. Accuracy of perceived speaking style in speech samples generated from different VAE z-vector centroids.

3.2. Controllability

In this experiment, we investigate the controllability of speaking style via the reference encoder. Figure 3 illustrates the VAE space for the systems augmented with conversational data from 4 speakers and 8 speakers, described in section 3.1. We achieve a reasonable separation of speaking styles, represented by clusters corresponding to individual speakers. Taking the system augmented with data from 4 speakers, we synthesize a set of samples conditioned on each of the VAE centroids, computed over the voice-converted training data from each speaker. We ask listeners to rate samples in terms of speaking style similarity against a natural reference produced by the target or supporting speakers. For each reference speaker, 30 listeners rated a total of 50 utterances in batches of 10 MUSHRA screens. For this evaluation, we asked listeners to ignore the identity of the voice and to focus solely on the similarity of the speaking style. To simplify our analysis, we took the highest rated sample for each submission to be the perceived speaking style. We then computed the accuracy for each centroid with respect to the reference speaker.

Results are summarized in Table 2 and indicate that listeners tend to perceive the correct speaking style. However, overall accuracy scores are still lower than expected and show some confusion across reference speakers. We hypothesize that this might be due to an inherent speaking style similarity across supporting speakers. Further work should validate these observations, evaluating systems with styles that are more perceptually different.

System	Naturalness	Speaker Sim.	Style Sim.
Recordings	64.69 ± 1.48	-	-
Neutral	56.71 ± 1.42	72.32 ± 1.48	40.21 ± 1.69
Data (1 hour)	57.82 ± 1.35	69.25 ± 1.45	55.41 ± 1.57
Data (3 hours)	57.58 ± 1.34	69.84 ± 1.46	56.68 ± 1.57
Data (6 hours)	58.19 ± 1.35	68.84 ± 1.48	56.03 ± 1.58
Data (8 hours)	56.55 ± 1.37	69.51 ± 1.48	55.49 ± 1.59
Source Speaker	-	28.33 ± 1.65	73.69 ± 1.40

Table 3. MUSHRA evaluation in terms of Naturalness, Speaker Similarity, and Speaking Style Similarity. Results indicate mean score with 95% confidence interval. "Data" systems denote TTS systems augmented with n hours of conversational data distributed equally across 4 supporting speakers. "Source Speaker" indicates a conversational TTS system trained on 8 hours from the source speaker.

3.3. Amount of supporting data

We investigate the amount of supporting data required for cross-speaker style transfer. As before, we train a TTS system on 10 hours of data in a neutral speaking style from the target speaker, termed *Neutral*. We also train a system on 8 hours of conversational data from the *Source speaker*. The cross-speaker style transfer systems are augmented with conversational data from 4 supporting speakers. We vary the total amount of available supporting conversational data, considering 8h, 6h, 3h, and 1h of data. The amount of data is distributed equally across the 4 conversational speakers. Therefore, we use 2 hours of data per speaker when considering a total of 8 hours of supporting data; and we take 15 minutes of data per speaker for the system augmented with 1 hour of data.

We follow the evaluation methodology described in section 3.1. **Results** are summarized in Table 3. We conduct two-sided t-tests on the MUSHRA scores with a Holm-Bonferroni correction on the three evaluations. For all listening tests, we observe no statistically significant differences between the augmented systems. For these systems, in terms of speaker similarity, we reduce the gap between source and target speaker systems by 93.3%. As before, augmented systems outperform the neutral system for style similarity. These results suggest that the proposed approach is effective with a reduced amount of supporting data.

3.4. Does it scale?

We validate our proposed approach on 14 different speakers, distributed across 7 dialects. We restrict the training data of each target speaker to a maximum of 10 hours of speech in a neutral reading style. For supporting data in a conversational speaking style, we consider 3 supporting speakers, each contributing with 1 hour of conversational data. *Augmented* systems are trained following the pipeline described in section 2. *Neutral* systems are trained using only the 10 hours of single-speaker neutral data. We evaluate systems in terms of speaker and style similarity. For speaker similarity, we follow the same paradigm as described before, although for simplicity we replace the synthetic source speaker sample with a vocoded sample. For style similarity, the reference is a sample from the source speaker in a conversational speaking style, matching the text of the synthetic utterances. We consider the neutral and augmented systems, and include as topline the corresponding voice converted sample. For each evaluation, 100 native speakers rated a set of 150 utterances, with each participant being assigned no more than 15 MUSHRA screens.

Lang	Gend	Speaker Sim				Style Sim			
		Src	Neut	Aug	*Rel*	VC	Neut	Aug	*Rel*
pt-PT	F	24.65	74.46	59.17	69.30%	68.89	46.85	65.22	83.35%
	M	15.52	78.94	74.10	92.37%	65.75	49.57	62.65	80.84%
pt-BR	F	26.17	79.32	67.81	78.34%	69.58	43.25	64.55	80.90%
	M	19.55	90.07	77.25	82.00%	70.12	56.41	63.98	55.22%
es-ES	F	35.67	76.53	71.94	88.77%	71.14	53.20	65.89	70.74%
	M	21.91	79.29	76.39	94.95%	73.41	57.75	67.89	64.75%
es-MX	F	47.66	65.67	65.58	99.50%	70.98	59.61	67.91	73.00%
	F	49.77	70.40	70.85	102.10%	68.95	63.07	65.61	43.20%
de-DE	F	26.40	77.30	73.11	91.77%	69.02	56.64	62.12	44.26%
	M	27.48	74.14	72.33	96.10%	64.68	55.73	57.20	16.42%
it-IT	F	25.67	69.42	67.25	95.00%	63.23	48.38	55.36	47.00%
	M	17.70	81.29	80.39	98.60%	71.22	54.19	59.20	29.42%
fr-CA	F	36.87	69.47	68.12	95.85%	65.87	55.57	61.26	55.24%
	F	38.10	68.75	67.31	95.30%	64.22	55.64	60.97	62.12%

Table 4. Cross-speaker style transfer for 14 speakers (9 female, 5 male). Results indicate mean MUSHRA score for speaker and style similarity. "Rel" indicates the relative position of the Augmented system ("Aug") to the remaining two competing systems: Neutral ("Neut"), source speaker ("Src"), or voice converted sample ("VC").

Results are presented in Table 4. In terms of speaker similarity, Augmented systems, on average, bridge the gap by 91.42% relative to the lower-anchor Source Speaker and the upper-anchor Neutral system ("Rel" column for speaker similarity in Table 4). We observe that only four systems bridge the gap by less than 90%. For style similarity, we consider the Neutral system to be the lower-anchor and the Voice Converted samples to be the upper-anchor. In this case, Augmented systems bridge the gap, on average, by 57.6% ("Rel" column for style similarity in Table 4). If we instead consider the Reference sample as our upper-anchor (assumed to be 100), then the Augmented systems bridge the gap, on average, by 18.5% over the Neutral samples. We additionally note that the difference between Augmented and Neutral systems is statistically significant at the level of $p < .01$ for all systems, except the German Male voice. Results indicate that our proposed approach is generally successful when transferring speaking style while preserving speaker identity. Typically, female target speakers perform better than male target speakers, which is likely due to a difference in speaking style between supporting and target speakers. Nonetheless, our results are extremely positive, in particular considering that we use only 3 hours of expressive data, with only 1 hour from the source speaker.

4. CONCLUSION AND FUTURE WORK

We addressed the problem of cross-speaker style transfer for TTS using data augmentation. Our approach uses a voice conversion model to generate high-quality data from a set of supporting expressive speakers. The voice converted samples are pooled with natural data from the target speaker to train a single-speaker multi-style TTS system. Results indicate that our proposed method works well when using a single or multiple supporting speakers, achieving good results with as little as 1 hour of expressive supporting data. Future work will focus on the development of the respective voice conversion and text-to-speech architectures. With respect to the TTS model architecture, we will investigate better disentanglement and stronger controllability of speaking style from the supporting conversational data. Additionally, we aim to scale the proposed methodology to more expressive speaking styles, such as emotions. Overall experimental results show that our proposed approach is efficient and scalable to multiple languages when transfer speaking style.

5. REFERENCES

[1] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*. IEEE, 2018, pp. 4779–4783.

[2] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[3] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "Speaker-dependent wavenet vocoder.," in *Proc. Interspeech*, 2017, pp. 1118–1122.

[4] Jaime Lorenzo-Trueba, Thomas Drugman, Javier Latorre, Thomas Merritt, Bartosz Putrycz, Roberto Barra-Chicote, Alexis Moinet, and Vatsal Aggarwal, "Towards achieving robust universal neural vocoding," in *Proc. Interspeech*, 2019.

[5] Keon Lee, Kyumin Park, and Daeyoung Kim, "STYLER: Style factor modeling with rapidity and robustness via speech decomposition for expressive and controllable neural text-to-speech," in *Proc. Interspeech*, 2021.

[6] Slava Shechtman, Raul Fernandez, Alexander Sorin, and David Haws, "Synthesis of expressive speaking styles with limited training data in a multi-speaker, prosody-controllable sequence-to-sequence architecture," *Proc. Interspeech*, pp. 4693–4697, 2021.

[7] Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proc. ICASSP*. IEEE, 2020, pp. 6189–6193.

[8] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *International conference on machine learning*, 2018, pp. 4693–4702.

[9] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*, 2018, pp. 5180–5189.

[10] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations, ICLR*, 2014.

[11] Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *Proc. ICASSP*. IEEE, 2019.

[12] Alexander Sorin, Slava Shechtman, and Ron Hoory, "Principal Style Components: Expressive style control and cross-speaker transfer in neural TTS.," in *Proc. Interspeech*, 2020.

[13] Matt Whitehill, Shuang Ma, Daniel McDuff, and Yale Song, "Multi-reference neural TTS stylization with adversarial cycle consistency," in *Proc. Interspeech*, 2020.

[14] Yanyao Bian, Changbin Chen, Yongguo Kang, and Zhenglin Pan, "Multi-reference Tacotron by intercross training for style disentangling, transfer and control in speech synthesis," *arXiv preprint arXiv:1904.02373*, 2019.

[15] Liumeng Xue, Shifeng Pan, Lei He, Lei Xie, and Frank K Soong, "Cycle consistent network for end-to-end style transfer tts training," *Neural Networks*, vol. 140, pp. 223–236, 2021.

[16] Xiaochun An, Frank K Soong, and Lei Xie, "Improving performance of seen and unseen speech style transfer in end-to-end neural TTS," *arXiv preprint arXiv:2106.10003*, 2021.

[17] Ajinkya Kulkarni, Vincent Colotte, and Denis Jouvet, "Improving transfer of expressivity for end-to-end multispeaker text-to-speech synthesis," *EUSIPCO 2021, EURASIP*, 2021.

[18] Young-Sun Joo, Hanbin Bae, Young-Ik Kim, Hoon-Young Cho, and Hong-Goo Kang, "Effective emotion transplantation in an end-to-end text-to-speech system," *IEEE Access*, vol. 8, pp. 161713–161719, 2020.

[19] Xiaochun An, Yuxuan Wang, Shan Yang, Zejun Ma, and Lei Xie, "Learning hierarchical representations for expressive speaking style in end-to-end speech synthesis," in *Proc. ASRU*. IEEE, 2019, pp. 184–191.

[20] Shifeng Pan and Lei He, "Cross-speaker style transfer with prosody bottleneck in neural speech synthesis," *Proc. Interspeech*, 2021.

[21] Goeric Huybrechts, Thomas Merritt, Giulia Comini, Bartek Perz, Raahil Shah, and Jaime Lorenzo-Trueba, "Low-resource expressive text-to-speech using data augmentation," in *Proc. ICASSP*. IEEE, 2021, pp. 6593–6597.

[22] Raahil Shah, Kamil Pokora, Abdelhamid Ezzerg, Viacheslav Klimkov, Goeric Huybrechts, Bartosz Putrycz, Daniel Korzekwa, and Thomas Merritt, "Non-autoregressive tts with explicit duration modelling for low-resource highly expressive speech," *Speech Synthesis Workshop (SSW11)*, 2021.

[23] Min-Jae Hwang, Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "TTS-by-TTS: TTS-driven data augmentation for fast and high-quality speech synthesis," in *Proc. ICASSP*. IEEE, 2021, pp. 6598–6602.

[24] Sri Karlapati, Alexis Moinet, Arnaud Joly, Viacheslav Klimkov, Daniel Sáez-Trigueros, and Thomas Drugman, "Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech," in *Proc. Interspeech*, 2020.

[25] Kaizhi Qian, Zeyu Jin, Mark Hasegawa-Johnson, and Gautham J Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in *Proc. ICASSP*. IEEE, 2020, pp. 6284–6288.

[26] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*, 2019, pp. 5210–5219.

[27] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. ICASSP*. IEEE, 2018, pp. 4879–4883.

[28] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *ICLR*, 2015.

[29] Yunlong Jiao, Adam Gabryś, Georgi Tinchev, Bartosz Putrycz, Daniel Korzekwa, and Viacheslav Klimkov, "Universal neural vocoding with parallel wavenet," in *Proc. ICASSP*, 2021.

[30] B Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

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