JASA Express Letters

MARCH 10 2026

Improving zero-shot style transfer text-to-speech by disentangled fine-grained style modeling

Eray Eren ; Qingju Liu ; Abeer Alwan ; Gaurav Bharaj

JASA Express Lett. 6, 034802 (2026)
https://doi.org/10.1121/10.0042974

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Improving zero-shot style transfer text-to-speech by disentangled fine-grained style modeling

Eray Eren,[1,a] Qingju Liu,[2,b] Abeer Alwan,[1,c] and Gaurav Bharaj[2,d]
[1]Department of Electrical and Computer Engineering, University of California, Los Angeles, California 90095, USA
[2]Flawless AI, Los Angeles, California 90401, USA

Abstract: Recent zero-shot style-transfer speech synthesis methods have shown promising results and addressed adaptation to unseen speaking styles. While most state-of-the-art methods generalize to new speakers and styles using large models or corpora, achieving similar generalization with a smaller model remains an open challenge. We propose a zero-shot method that uses the small GenerSpeech backbone plus a fine-grained style encoder. To disentangle speakers, global/fine-grained styles, and content embeddings, we introduce a mutual-information minimization loss. To further disentangle style from speaker and boost style embedding diversity, we introduce a maximum-mean-discrepancy-guided cycle consistency loss. Experimental results show the proposed method outperforms baseline zero-shot style-transfer methods (GenerSpeech, YourTTS, VALL-E-X) with a relative average style preference improvement of 31% and a 3.64 prosody prosody similarity mean opinion score on VCTK.

[Editor: Douglas D. O'Shaughnessy] https://doi.org/10.1121/10.0042974

Received: 28 October 2025 **Accepted:** 24 February 2026 **Published Online:** 10 March 2026

1. Introduction

Human-level performance has been achieved by single-speaker modern text-to-speech (TTS) systems (Kim *et al.*, 2021; Tan *et al.*, 2024). Yet, challenges remain in model-efficient zero-shot style-transferable TTS. First, parameter-inefficient large models consume extensive resources (Du *et al.*, 2024; Lajszczak *et al.*, 2024; Le *et al.*, 2023; Wang *et al.*, 2025; Zhang *et al.*, 2023). Second, a fine-grained style transfer from a reference speech remains a challenge. This may cause style mismatch, such as differences in pitch or duration statistics.

Several studies have attempted to address style transfer challenges. For example, in Chen and Rudnicky (2022), a wav2vec2-based (Baevski *et al.*, 2020) style network is used to extract temporal style embeddings and a global speaker embedding from a reference speech signal. In Li *et al.* (2025), a global style embedding is extracted to control predictions of frame-level pitch & energy. The method in Min *et al.* (2021) uses style-adaptive layer normalization to condition both the text encoder and Mel-decoder with the extracted global style. The VITS-based (Kim *et al.*, 2021), cross-lingual method (Casanova *et al.*, 2022) uses additional language and speaker embeddings, and a speaker-consistency loss for improving zero-shot adaptation performance.

In Guo *et al.* (2023), an emerging prompt-guiding TTS system uses text prompts to explicitly describe speech styles. However, this requires style-rich text prompts to explicitly label the speech style and voice variability. The subsequent research (Leng *et al.*, 2024) mitigates the need for human-labeled style attributes. It uses automatic labeling of prosodic or style features, together with a large language model to synthesize text prompts.

In addition to the above methods focused on global style encodings, other studies have explored hierarchical or fine-grained style encoding for better style granularity. In GenerSpeech (Huang *et al.*, 2022), besides the global speaker and emotion embeddings, a multi-level style encoder is also utilized to extract local style embeddings. NaturalSpeech 3 (Ju *et al.*, 2024) employs factorized vector quantization to obtain time-varying prosody (style) disentangled from other speech attributes (identity, content, acoustic details), with the help of information bottleneck, various supervised losses, and adversarial training.

Despite the aforementioned methods with explicit style modeling, the prevailing trend in TTS is to exploit large models (Du *et al.*, 2024; Lajszczak *et al.*, 2024; Wang *et al.*, 2025; Zhang *et al.*, 2023) with implicit style modeling, where styles are embedded in context or discrete neural acoustic codecs (Défossez *et al.*, 2023). However, these models require not only a large parameter space on the order of hundreds of millions of parameters, but also very large training datasets

[a]Email: erayeren@ucla.edu
[b]Email: qingju.liu@flawlessai.com
[c]Corresponding author: alwan@ee.ucla.edu
[d]Email: gaurav.bharaj@gmail.com

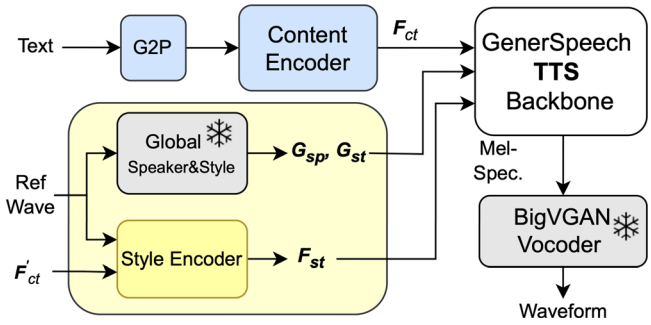

Fig. 1. Overview of the proposed model architecture, featuring pitch & energy predictors embedded within the backbone. These predictors utilize the fine-grained style $\mathbf{F}_{st}$ vector from our proposed style encoder, and the additional global speaker $\mathbf{G}_{sp}$ and global style $\mathbf{G}_{st}$ vectors. Style encoder input $\mathbf{F}'_{ct}$ is the summation of length-regulated (LR) $\mathbf{G}_{sp}$, $\mathbf{F}_{st}$, and content $\mathbf{F}_{ct}$ (shown in Fig. 2). The snowflake indicates pretrained and frozen models.

containing tens of thousands of hours of speech signals. Smaller models (e.g., Casanova *et al.*, 2022; Huang *et al.*, 2022) have not fully addressed style transfer with unseen speakers or styles, and some methods (Huang *et al.*, 2022) require external automatic speech recognition models or aligners at inference time.

To address these issues, we propose a small-size TTS model (48 M parameters) with fine-grained style control. We propose a novel style encoder that does not require external phoneme or word alignment information at inference time. The proposed style encoder is accompanied by novel loss functions achieving twofold disentanglement. First, speaker, style, and content embeddings are disentangled using a specific comprehensive mutual information minimization (MIM) loss. Second, the global style and speaker embeddings are further disentangled with a maximum mean discrepancy (MMD)-guided cycle consistency loss. Experiments show that our model outperforms baseline zero-shot style transfer models in terms of prosody similarity mean opinion score (PMOS) as well as A-B-X (ABX) style preference tests, where A and B are synthetic outputs (ours vs baseline) and X is the ground-truth reference style. We provide audio samples at https://profiterole1107.github.io/styletransferdisentanglement/.

2. Methods

The proposed model architecture is shown in Fig. 1, where we use GenerSpeech (Huang *et al.*, 2022) as the backbone. The TTS backbone utilizes fine-grained style ($\mathbf{F}_{st}$), global speaker ($\mathbf{G}_{sp}$), and global style ($\mathbf{G}_{st}$) embeddings to produce a Mel-spectrogram from text using style-variant and style-invariant branches (pitch & energy predictors) similar to Huang *et al.* (2022). Given a specific text, the non-autoregressive transformer-based content encoder (Ren *et al.*, 2021) encodes phoneme embeddings from the grapheme-to-phoneme model (Park, 2019). The encoded content $\mathbf{F}_{ct}$ is fed into mixed-style layer normalization to make it invariant to the global style and speaker information similar to Huang *et al.* (2022). The style-variant predictors are conditioned on $\mathbf{G}_{sp}$, $\mathbf{G}_{st}$, and $\mathbf{F}_{st}$. GenerSpeech employs hierarchical phoneme-, word- and utterance-level style encoders. These encoders require explicit alignment information at inference time. We propose a novel style encoder (Fig. 2) that does not rely on alignment information. Moreover, we designed new training objectives with twofold disentanglement, described in Sec. 2.1.

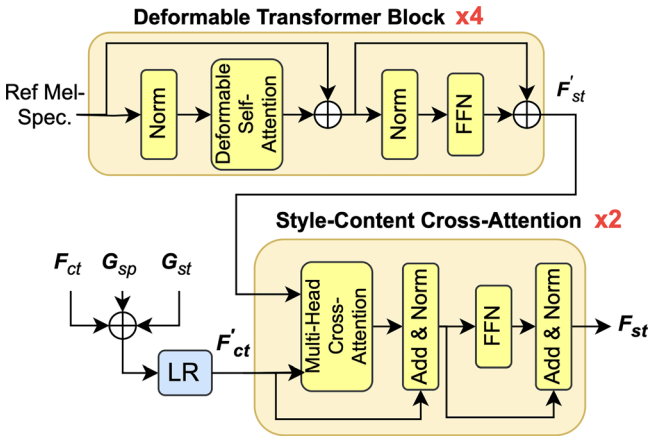

Fig. 2. Proposed style encoder with 4 deformable transformer blocks and 2 cross-attention modules.

2.1 Style modeling

The proposed style encoder consists of deformable transformer blocks (DTB) and style-content cross-attention blocks (SC-CABs). The deformable self-attention blocks have been shown to learn multi-granular information effectively (Chen *et al.*, 2023), thus avoiding dependency on hierarchical encoders and external aligners in Huang *et al.* (2022). The DTB takes advantage of deformable self-attention in Chen *et al.* (2023), which is one type of multi-head attention with learnable attention window offsets and attention window sizes. A stack of these blocks can learn fine-grained multi-level style information. This alleviates the requirement for a defined set of phoneme and word boundaries as in Huang *et al.* (2022), and hence an external aligner (or automatic speech recognition) is not required at inference time.

The stacked DTB outputs $\mathbf{F}'_{\mathbf{st}}$ spans the same temporal length as the reference audio. To align with the length-regulated content $\mathbf{F}'_{\mathbf{ct}}$, we use a stack of SC-CABs after the DTB to obtain the fine-grained style $\mathbf{F}_{\mathbf{st}}$, which are then used for conditioning the style-variant pitch & energy predictors as well as the Mel-decoder and post-net.

Although the fine-grained and global style-related vectors incorporate prosody, they may contain other information, such as speaker or content. Therefore, similar to Ju *et al.* (2024), we disentangle styles from speakers and content in an unsupervised manner. However, our approach is different from Ju *et al.* (2024) and uses two different strategies as described below. The style information is defined as the remaining component after speaker and content disentanglement.

Mutual Information Minimization (MIM): In Sigurgeirsson and King (2023), the authors hypothesize that some style transfer models do not learn a transferable representation of prosody, but rather an utterance-level representation that depends on both the reference speaker and reference text. This dependency significantly degrades performance for unseen (speaker, text) pairs at inference time. Hence, we propose to disentangle speaker, style, and content.

Mutual information measures the dependency between two random variables. To minimize the dependency between speaker ($\mathbf{G}_{\mathbf{sp}}$) and content ($\mathbf{F}_{\mathbf{ct}}$) encoders, we utilize the variational contrastive log-ratio upper bound (vCLUB) (Cheng *et al.*, 2020). Since the true conditional distribution is unknown, vCLUB uses a variational approximation q_θ to define the upper bound as follows:

$$I(\mathbf{G_{sp}};\mathbf{F_{ct}}) = \mathbb{E}_{p(G_{sp};F_{ct})}\left[\log(q_\theta(\mathbf{G_{sp}}|\mathbf{F_{ct}}))\right] - \mathbb{E}_{p(\mathbf{G_{sp}})}\mathbb{E}_{p(\mathbf{F_{ct}})}\left[\log q_\theta(\mathbf{G_{sp}}|\mathbf{F_{ct}})\right], \tag{1}$$

where the first term evaluates the log-likelihood of positive (dependent) pairs, and the second term evaluates negative (independent) pairs. To ensure a reliable upper bound, a neural network is used to estimate Q (parameterized by θ) to approximate the true distribution by minimizing the Kullback-Leibler divergence as follows:

$$\min_{Q} KL(p(\mathbf{G_{sp}}|\mathbf{F_{ct}}) \parallel q_\theta(\mathbf{G_{sp}}|\mathbf{F_{ct}})). \tag{2}$$

Given Q, we calculate the empirical vCLUB (approximation of 1) on a mini-batch (a subset of training samples) of size N and sequence length T as follows:

$$\hat{I}(\mathbf{G_{sp}};\mathbf{F_{ct}}) = \frac{1}{N^2T}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\left[\log q_\theta(\mathbf{G}^i_{\mathbf{sp}}|\mathbf{F}^{i,t}_{\mathbf{ct}}) - \log q_\theta(\mathbf{G}^j_{\mathbf{sp}}|\mathbf{F}^{i,t}_{\mathbf{ct}})\right]. \tag{3}$$

For each step, we first update Q [Eq. (2)] to learn the dependency, and then update $\mathbf{G_{sp}}$ and $\mathbf{F_{ct}}$ [Eq. (3)] to remove it (minimize mutual information). To perform disentanglement, the same strategy is applied between the following pairs of embeddings: $\mathbf{G_{sp}}$ vs $\mathbf{G_{st}}$, $\mathbf{G_{sp}}$ vs $\mathbf{F_{st}}$, $\mathbf{G_{st}}$ vs $\mathbf{F_{ct}}$.

Cycle-Consistency: The cycle consistency loss (An *et al.*, 2022; Jo *et al.*, 2023; Xue *et al.*, 2021) has been used for disentangling speaker and style attributes for expressive speech synthesis systems with style transfer capabilities. Conceptually, cycle consistency ensures that translating a sample to a target domain and then back to the source domain yields the original sample, promoting structure preservation (Xue *et al.*, 2021). Our experimental findings reveal that this cyclic loss leads to decreased variance for the predicted style embeddings. That is, the model's style predictions become closer to each other, which is not beneficial for diverse speech generation. To increase the diversity of the predicted speaker and style embeddings, we propose a MMD (Gretton *et al.*, 2012) with a Gaussian kernel to guide the cyclic-consistency loss.

The aim of cyclic losses is to create new speaker and style pairings, and cycle back to the same style and speaker pairing. Then, we extract global speaker and style vectors from Mel-spectrograms using these pairings, and attempt to make them closer to the corresponding speaker or style vectors.

To achieve this, first, we randomly sample an utterance with a text ($\mathbf{Text}^1$), a global speaker embedding ($\mathbf{G}^1_{sp}$), and a global style embedding ($\mathbf{G}^1_{st}$). We randomly sample another utterance from another speaker that has an embedding $\mathbf{G}^2_{sp}$ with a global style embedding $\mathbf{G}^2_{st}$. As shown in Fig. 3, we synthesize the Mel-spectrogram $\widehat{\mathbf{Mel}}^{st^1}_{sp^2}$ using $\mathbf{Text}^1$, $\mathbf{G}^1_{st}$, and $\mathbf{G}^2_{sp}$. We then extract the global style $\hat{\mathbf{G}}^1_{st}$ and speaker embeddings $\hat{\mathbf{G}}^2_{sp}$ from that synthesized Mel-spectrogram. Using the extracted style

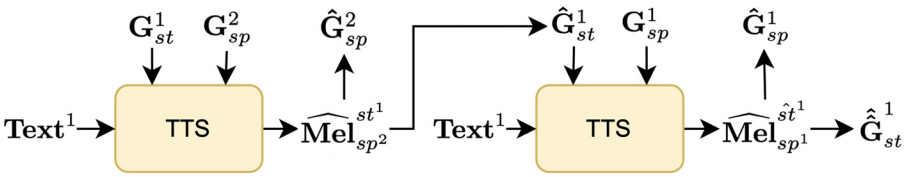

Fig. 3. Cycle consistency illustration.

embedding $\hat{\mathbf{G}}^1_{st}$ and the original speaker embedding $\mathbf{G}^1_{sp}$, we synthesize $\widehat{\mathbf{Mel}}^{\hat{st}^1}_{sp^1}$ with $\mathbf{Text}^1$. Similarly, we extract the global style $\hat{\hat{\mathbf{G}}}_{st}1$ and speaker embedding $\hat{\mathbf{G}}^1_{sp}$ from $\widehat{\mathbf{Mel}}^{\hat{st}^1}_{sp^1}$. The cycle consistency loss L_{cycle} is then calculated as follows:

$$\begin{aligned} L_{cycle} = {} & L_2(\mathbf{Mel}^{st^1}_{sp^1}, \widehat{\mathbf{Mel}}^{\hat{st}^1}_{sp^1}) + L_2(\mathbf{G}^1_{st}, \hat{\mathbf{G}}^1_{st}) + L_2(\mathbf{G}^1_{st}, \hat{\hat{\mathbf{G}}}_{st}1) + L_2(\mathbf{G}^1_{sp}, \hat{\mathbf{G}}^1_{sp}) + L_2(\mathbf{G}^2_{sp}, \hat{\mathbf{G}}^2_{sp}) \\ & + \lambda L_{MMD}(\mathbf{G}^1_{st}, \mathbf{G}^2_{st}) + \lambda L_{MMD}(\mathbf{G}^1_{st}, \mathbf{G}^{avg}_{st}), \end{aligned} \tag{4}$$

where $\mathbf{Mel}^{st^1}_{sp^1}$ indicates the ground-truth Mel-spectrogram that belongs to the style 1 and speaker 1, and $\mathbf{G}^{avg}_{st}$ denotes the average style vector of the mini-batch. *L2* and L_{MMD} indicate the mean square error and biased MMD (Gretton *et al.*, 2012), respectively. Lambda (λ) denotes the weights of the L_{MMD} losses.

The losses of Eq. (4) are for ensuring resynthesis quality (the *L2* term of Mel-spectrograms), disentangling global speaker/style embeddings (the *L2* terms of different G_{sp} and G_{st} vectors), and boosting global style diversity (L_{MMD}). Cycle and MIM losses are added to the original GenerSpeech losses (Mel-spectrogram, duration, pitch & energy).

3. Experiments

3.1 Experimental settings

The VCTK dataset (Veaux *et al.*, 2016), with 108 speakers and 44 h in duration, was used for the experiments. Similar to Casanova *et al.* (2022), we excluded 10 speakers from training to use them for testing zero-shot speaker adaptation and style transfer. We employed three high-performing baselines: YourTTS (Casanova *et al.*, 2022), VALL-E-X (Plachtaa, 2023; Zhang *et al.*, 2023), and GenerSpeech (Huang *et al.*, 2022). For a fair comparison, we used the same pretrained global embeddings in the GenerSpeech model (Huang *et al.*, 2022) with the same train-validation-test split of the dataset, and we also used the same vocoder (Lee *et al.*, 2023).

Recordings were downsampled to 16 kHz. A window of 64 ms with 16 ms hop size was used for extracting 80-dimensional Mel-spectrograms. We used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a batch size of 64. The BigVGAN vocoder (Lee *et al.*, 2023) reconstructs raw waveforms directly from Mel-spectrograms. On the other hand, YourTTS learns hidden acoustic features and converts them into waveforms using a HifiGAN-based vocoder, which is very similar to BigVGAN vocoder in terms of model structure. VALL-E-X also makes use of hidden acoustic features, but it has a convolutional network-based VQ-VAE (Défossez *et al.*, 2023) decoder to generate the waveforms. The pitch is extracted/measured using standard, auto-correlation-based F0 extraction algorithms in Praat (Boersma, 1993). During inference, Praat-based F0 values serve as ground-truth measurements; and during training, these F0 values are used as training targets of the pitch predictor network. The network predicts auxiliary F0 values for more accurate Mel-spectrogram predictions.

Lambda (λ) for the mean discrepancy was -0.1 with a Gaussian kernel to guide the cyclic losses. In total, there were 48×10^6 (M) trainable parameters. Our model was trained on a single NVIDIA A10G GPU for 300 K training steps. The duration, pitch, and energy predictors consist of two Conv1D layers (kernel size of 3, and 256-D input and output channel dimensions) followed by ReLU, layer normalization, dropout (dropout rate of 0.1) and a linear output layer (256-D). The global speaker embedding $\mathbf{G_{sp}}$ and global style vector $\mathbf{G_{st}}$ were extracted with pretrained models from Wan *et al.* (2018). For fine-tuning purposes, both $\mathbf{G_{sp}}$ and $\mathbf{G_{st}}$ are further projected with an additional 256-D linear mapping layer. For the proposed style encoder, we used a stack of 4 DTBs with 4 attention heads and a 512-D embedding size, and 2 linear layers with GeLU activation and dropout layers (rate of 0.1) for the feed-forward network in each DTB. The number of SC-CAB blocks is 2 with 2 attention heads and a 2048-D embedding size. The feed-forward network of each SC-CAB was similar to the feed-forward network of DTB.

For each test sample $x(n)$, we randomly chose a sample $r(n)$, as a reference sample, from the same speaker. We then ran TTS by transferring the style from $r(n)$ using the content from $x(n)$, and the result is $\hat{x}(n)$. We employed a pretrained BigVGAN vocoder to convert the synthesized Mel-spectrogram into a time-domain waveform. To account for errors caused by the vocoder, we also applied the vocoder to the Mel-spectrogram of the original $r(n)$, denoted as vocoded ground-truth or GT-V.

We conducted a subjective listening test with Amazon Mechanical Turk, where we recruited 25 candidates to assess different methods using two evaluation metrics: a quality mean opinion score (MOS) and a PMOS. For PMOS,

subjects were instructed to give a similarity score between the reference signal $r(n)$ and the synthesized signal $\hat{x}(n)$ in terms of manner of speaking, voice tone, and emphasis. In addition, an ABX preference test for our model and different baselines were carried out to evaluate the style transfer preference with similar instructions. Each listener evaluated 25 groups of samples randomly chosen from the test dataset for each baseline. The MOS, PMOS, and ABX scores were then averaged.

For objective evaluations, we considered UTMOS (Nakata *et al.*, 2024) for measuring the perceptual quality of synthesized speech. Additionally, we used AutoPCP (Seamless, 2023) to objectively evaluate the prosody similarity of the reference and synthesized speech.

3.2 Experimental results

Table 1 shows that listeners gave the highest PMOS scores to our model, significantly outperforming the baselines. This shows the effectiveness of the proposed style encoder together with the disentanglement boosted by MIM and mean discrepancy-guided cycle consistency. Consistent improvements extend to the AutoPCP evaluations. The VALL-E-X model has the highest MOS and UTMOS scores, which is possibly due to its much larger model size and training set. Our model's MOS and UTMOS scores are similar to GenerSpeech, with the advantage of not relying on an external aligner at inference time.

Moreover, our method is capable of extracting and transferring style in a manner that more closely aligns with human perception, as shown by the ABX style preference listening tests in Table 2. On average, a 7-point score of 0.996 was observed with a relative average style preference improvement of 31% (compared to baselines).

The proposed method generates phoneme-level pitch and duration statistics that are more similar to the ground-truth compared to the other baselines, with the exception of YourTTS phoneme-level pitch mean values. This is shown in Table 3, where we calculated statistics (mean, standard deviation, skewness, kurtosis) of phoneme-level pitch and duration. These analyses demonstrate the capacity of a method to preserve prosody and style (Shen *et al.*, 2024). This is done by first applying an external aligner (McAuliffe *et al.*, 2017) over a synthesized signal and the reference signal, and then extracting the average pitch and duration at each phoneme segment. The aforementioned statistics are then calculated over the absolute difference of these two phoneme-level features.

To quantify cross-domain performance, we use the LibriTTS train set (clean) to train our model, and report the results in Table 3. Similar to in-domain performance, the proposed model can capture the prosody/style better than the baselines, indicating cross-domain robustness.

3.3 Ablation study

To investigate the contributions from different modules, we carried out an ablation study as shown in Table 4. Two models were trained independently where each model represents the original model but without a particular loss: (1) the cycle consistency loss and (2) the MIM loss. In addition to UTMOS and AutoPCP, we also assessed the mean absolute error (MAE) over phoneme-level duration, and the root mean square error (RMSE) over phoneme-level pitch with respect to the ground-truth. The results show that removing the cycle loss or MIM degrades UTMOS and AutoPCP. Furthermore, removal of cycle loss or MIM significantly makes the F0 estimation inferior, although there is a slight improvement in duration estimation.

We quantify the between-vector disentanglement with the Gaussian radial-basis-function centered kernel alignment (RBF-CKA) (Kornblith *et al.*, 2019) to capture the nonlinear dependence. RBF-CKA is invariant to orthogonal transforms (and to translations); however, it is not intrinsically invariant to isotropic scaling. Therefore, to ensure that scores are directly comparable across methods and datasets, we first center and scale each vector to zero mean and unit variance per feature, and set the RBF bandwidth using the median pairwise distance heuristic computed on the vectors. Table 5 shows that removing MIM or the cycle loss increases nonlinear dependence, indicating these losses promote disentanglement. Note that $\mathbf{G_{ct}}$ is computed as the temporal average of the content representation $\mathbf{F_{ct}}$.

Table 1. Subjective evaluations in terms of MOS and PMOS for unseen speakers; objective evaluations include UTMOS, AutoPCP with zero-shot adaptation. Up arrows indicate performance improvement. The ground-truth is vocoded (GT-V). The second column shows the number of model parameters.

Model	Size	MOS ↑	PMOS ↑	UTMOS ↑	AutoPCP ↑
GT-V	—	4.15	—	3.29	—
YourTTS	87M	3.29	3.38	2.61	1.95
VALL-E-X	300M	3.66[a]	3.56	3.06[a]	2.09
GenerSpeech	52M	3.58[b]	3.53	2.96[b]	2.11
Proposed model	48M	3.55	3.64[a]	2.97	2.24[a]

[a]Best values.

[b]Statistically insignificant ($p > 0.05$) results compared to the proposed model.

Table 2. Subjective evaluations of zero-shot style transfer capacity for unseen speakers in terms of ABX preference test scores, where the participants scored baseline models (A) against our model (B) with a 7-point scale ranging from –3 to 3. A score of 0 or "Neutral" indicates the same style preference. A positive score indicates an overall preference for our model style (B).

Model	7-point score	A (%)	Neutral	B (%)
YourTTS	1.08	31	10	59
VALL-E-X	1.25	26	11	63
GenerSpeech	0.75	23	25	52

Table 3. Statistical analyses of differences in phoneme-level pitch/duration between the reference, baselines, and our method. All models are tested on VCTK; left block: models trained on VCTK (except VALL-E-X), right block: models trained on LibriTTS (except VALL-E-X). Down arrows indicate decrease.

	Train: VCTK				Train: LibriTTS			
Model	Mean ↓	SD ↓	Skewness ↓	Kurtosis ↓	Mean ↓	SD ↓	Skewness ↓	Kurtosis ↓
YourTTS	11.98[a]/0.009	4.855/0.011	0.235/0.655	0.236/2.841	15.04/0.009	4.171/0.008[a]	0.164/0.694	0.199/4.250
VALL-E-X	24.19/0.009	4.868/0.011	0.226/0.718	0.220/2.955	24.19/0.009	4.868/0.011	0.226/0.718	0.220/2.955
GenerSpeech	14.41/0.007	4.625/0.009[a]	0.231/0.649	0.248/2.613	12.78/0.008	3.348/0.009	0.162/0.522	0.174[a]/2.706
Proposed model	12.89/0.004[a]	4.196[a]/0.010	0.218[a]/0.621[a]	0.215[a]/2.511[a]	12.75[a,b]/0.006	3.129/0.008[a,b]	0.150[a]/0.470[a]	0.179/2.463[a]

[a]Best values.
[b]Statistically insignificant ($p > 0.05$) results compared to the proposed model.

Table 4. Ablations with two models, which separately removes the cycle consistency loss and MIM. UTMOS, AutoPCP, pitch RMSE (RMSE F0), and duration MAE are reported using the ground-truth speech as the reference speech. Up arrows indicate improvement. Down arrows indicate decrease.

Model	UTMOS ↑	AutoPCP ↑	RMSE F0 ↓	Duration MAE ↓
Proposed model	3.04[a]	2.33[a]	26.9[a]	0.049
Cycle loss	2.95	2.11	31.2	0.045
MIM	2.93	2.08	29.2	0.043[a]

[a]Best values.

Table 5. Ablations with two models, which separately removes the cycle consistency loss and MIM. Gaussian RBF-CKA scores are reported for quantifying disentanglement between global speaker ($\mathbf{G_{sp}}$), style ($\mathbf{G_{st}}$), and content ($\mathbf{G_{ct}}$) vectors. Down arrows indicate decrease.

Model	$\mathbf{G_{sp}} - \mathbf{G_{st}}$ ↓	$\mathbf{G_{st}} - \mathbf{G_{ct}}$ ↓	$\mathbf{G_{sp}} - \mathbf{G_{ct}}$ ↓
Proposed model	0.16[a]	0.15[a]	0.14[a]
Cycle loss	0.32	0.23	0.22
MIM	0.38	0.24	0.21

[a]Best values.

4. Conclusion

We introduce a novel zero-shot style-transferable TTS system with fine-grained style encoding. The style encoder does not require alignment (unlike our backbone) at inference. Mutual information minimization, together with a MMD-guided cycle consistency loss, are applied for the disentanglement of style, speaker, and content. Despite the additional computational demands posed by cycle consistency during training, no computation is required for MIM or cycle consistency during inference. Our experimental results show that the proposed method outperforms several baselines for both objective and subjective evaluations in terms of style similarity with a relatively small model footprint (48×10^6 parameters). Specifically, our method yielded a PMOS score of 3.64, and with a relative average style preference improvement of 31%.

Author Declarations

Conflict of Interest

E.E. was an intern at Flawless AI during the conduct of this work. Q.L. and G.B. were employees of Flawless AI during the conduct of this work. The authors have no other conflicts to disclose.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. A representative subset of the listening samples used in this study is publicly available at https://profiterole1107.github.io/styletransferdisentanglement/.

References

An, X., Soong, F. K., and Xie, L. (**2022**). "Disentangling style and speaker attributes for TTS style transfer," IEEE/ACM Trans. Audio Speech Lang. Process. **30**, 646–658.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (**2020**). "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, Vol. 33, pp. 12449–12460.

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, Amsterdam, Vol. 17, pp. 97–110.

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

Chen, L., and Rudnicky, A. (**2022**). "Fine-grained style control in transformer-based text-to-speech synthesis," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 7907–7911.

Chen, W., Xing, X., Xu, X., Pang, J., and Du, L. (**2023**). "DST: Deformable speech transformer for emotion recognition," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5.

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. (**2020**). "CLUB: A contrastive log-ratio upper bound of mutual information," in *International Conference on Machine Learning* (PMLR), pp. 1779–1788.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. (**2023**). "High fidelity neural audio compression," in *Transactions on Machine Learning Research*.

Du, Z., Wang, Y., Chen, Q., Shi, X., Lv, X., Zhao, T., Gao, Z., Yang, Y., Gao, C., Wang, H., Yu, F., Liu, H., Sheng, Z., Gu, Y., Deng, C., Wang, W., Zhang, S., Yan, Z., and Zhou, J. (**2024**). "Cosyvoice 2: Scalable streaming speech synthesis with large language models," arXiv:2412.10117.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. J. (**2012**). "A kernel two-sample test," J. Mach. Learn. Res. **13**(1), 723–773.

Guo, Z., Leng, Y., Wu, Y., Zhao, S., and Tan, X. (**2023**). "PromptTTS: Controllable text-to-speech with text descriptions," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5.

Huang, R., Ren, Y., Liu, J., Cui, C., and Zhao, Z. (**2022**). "Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech," in *Advances in Neural Information Processing Systems*, Vol. 35, pp. 10970–10983.

Jo, S., Lee, Y., Shin, Y., Hwang, Y., and Kim, T. (**2023**). "Cross-speaker emotion transfer by manipulating speech style latents," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5.

Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, E., Leng, Y., Song, K., and Tang, S. (**2024**). "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," in *International Conference on Machine Learning*, pp. 22605–22623.

Kim, J., Kong, J., and Son, J. (**2021**). "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning* (PMLR), pp. 5530–5540.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (**2019**). "Similarity of neural network representations revisited," in *International Conference on Machine Learning* (PMLR), pp. 3519–3529.

Lajszczak, M., Cámbara, G., Li, Y., Beyhan, F., Van Korlaar, A., Yang, F., Joly, A., Martőn-Cortinas, Á., Abbas, A., and Michalski, A. (**2024**). "Base TTS: Lessons from building a billion-parameter text-to-speech model on 100k hours of data," arXiv:2402.08093.

Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., and Hsu, W. (**2023**). "Voicebox: Text-guided multilingual universal speech generation at scale," in *Advances in Neural Information Processing Systems*, Vol. 36, pp. 14005–14034.

Lee, S., Ping, W., Ginsburg, B., Catanzaro, B., and Yoon, S. (**2023**). "BigVGAN: A universal neural vocoder with large-scale training," in *International Conference on Learning Representations*.

Leng, Y., Guo, Z., Shen, K., Tan, X., Ju, Z., Liu, Y., Liu, Y., Yang, D., Zhang, L., Song, K., He, L., Li, X., Zhao, S., Qin, T., and Bian, J. (**2024**). "PromptTTS 2: Describing and generating voices with text prompt," in *International Conference on Learning Representations*.

Li, Y. A., Han, C., and Mesgarani, N. (**2025**). "StyleTTS: A style-based generative model for natural and diverse text-to-speech synthesis," IEEE J. Sel. Top. Signal Process. **19**(1), 283–296.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (**2017**). "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Interspeech*, pp. 498–502.

Min, D., Lee, D. B., Yang, E., and Hwang, S. J. (**2021**). "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *International Conference on Machine Learning*, pp. 7748–7759.

Nakata, W., Yamauchi, K., Yang, D., Hyodo, H., and Saito, Y. (**2024**). "UTDUSS: Utokyo-Sarulab system for Interspeech2024 speech processing using discrete speech unit challenge," arXiv:2403.13720.

Park, K., and Kim, J. (**2019**). "g2pe," https://github.com/Kyubyong/g2p.

Plachtaa (**2023**). "VALL-E-X," https://github.com/Plachtaa/VALL-E-X.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T. (**2021**). "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*.

Seamless (**2023**). "Seamless: Multilingual expressive and streaming speech translation," arXiv:2312.05187.

Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., Qin, T., Zhao, S., and Bian, J. (**2024**). "NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *International Conference on Learning Representations*.

Sigurgeirsson, A. T., and King, S. (**2023**). "Do prosody transfer models transfer prosody*f*," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5.

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Soong, F. K., Qin, T., Zhao, S., and Liu, T. (**2024**). "NaturalSpeech: End-to-end text to speech synthesis with human-level quality," IEEE Trans. Pattern Anal. Mach. Intell. **46**(6), 4234–4245.

Veaux, C., Yamagishi, J., and MacDonald, K. (**2016**). "Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," https://datashare.ed.ac.uk/handle/10283/2651.

Wan, L., Wang, Q., Papir, A., and Lopez-Moreno, I. (**2018**). "Generalized end-to-end loss for speaker verification," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 4879–4883.

Wang, Y., Zhan, H., Liu, L., Zeng, R., Guo, H., Zheng, J., Zhang, Q., Zhang, S., and Wu, Z. (**2025**). "MaskGCT: Zero-shot text-to-speech with masked generative codec transformer," in *International Conference on Learning Representations*.

Xue, L., Pan, S., He, L., Xie, L., and Soong, F. K. (**2021**). "Cycle consistent network for end-to-end style transfer TTS training," in *Neural Networks*, Vol. 140, pp. 223–236.

Zhang, Z., Zhou, L., Wang, C., Chen, S., Wu, Y., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., and Wei, F. (**2023**). "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," arXiv:2303.03926.