INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

A Universal Multi-Speaker Multi-Style Text-to-Speech via Disentangled Representation Learning based on Rényi Divergence Minimization

Dipjyoti Paul[1]*, Sankar Mukherjee*[2]*, Yannis Pantazis*[3] *and Yannis Stylianou*[1]

[1]Computer Science Department, University of Crete, Greece
[2]Istituto Italiano di Tecnologia, Italy
[3]IACM, Foundation for Research and Technology - Hellas, Greece

dipjyotipaul@csd.uoc.gr, sankar1535@gmail.com, pantazis@iacm.forth.gr, yannis@csd.uoc.gr

Abstract

In this paper, we present a universal multi-speaker, multi-style Text-to-Speech (TTS) synthesis system which is able to generate speech from text with speaker characteristics and speaking style similar to a given reference signal. Training is conducted on non-parallel data and generates voices in an unsupervised manner, i.e., neither style annotation nor speaker label are required. To avoid leaking content information into the style embeddings (referred to as "content leakage") and leaking speaker information into style embeddings (referred to as "style leakage") we suggest a novel **R**ényi **D**ivergence based **D**isentangled **R**epresentation framework through adversarial learning. Similar to mutual information minimization, the proposed approach explicitly estimates via a variational formula and then minimizes the Rényi divergence between the joint distribution and the product of marginals for the content-style and style-speaker pairs. By doing so, content, style and speaker spaces become representative and (ideally) independent of each other. Our proposed system greatly reduces content leakage by improving the word error rate by approximately 17-19% relative to the baseline system. In MOS-speech-quality, the proposed algorithm achieves an improvement of about 16-20% whereas MOS-style-similarly boost up 15% relative performance.

Index Terms: Unsupervised style-content-speaker separation, Rényi divergence minimization, representation disentanglement, controllable speech synthesis

1. Introduction

Speech synthesis which attracts a lot of attention in communication and voice interaction systems aims to synthesize intelligible and high quality speech signals which are indistinguishable from human recordings. The realization of a spoken utterance can be categorized into three principal components: the content, the speaker and the style component. The content component refers to the linguistic content of speech (what). The speaker/voice characteristics are attributed to the speaker component (who). While the definition of style component associates with the variation of pitch, loudness and prosody (how). Style generally covers all aspects of speech that does not contribute to content information or the identification of the speaker.

Recently, the superiority of deep neural network (DNN) based speech synthesis surpassed the conventional speech synthesis models [1, 2, 3, 4, 5, 6]. Given a sufficient amount of training data, such TTS systems are capable of producing speech with superior quality, particularly for single-speaker synthesis. However, generated speech usually tends to be neutral and less expressive. Synthetic speech expressivity is also restricted by the fact that collecting labelled speech data especially information describing prosody is cumbersome due to concerns on cost, complexity and privacy making unsupervised representation learning immensely popular.

Relating to expressive TTS, previous works aimed to transfer the style factor of a reference speech into the given text without prosody labels [7, 8]. In Global Style Token (GST), the reference speech is encoded into a fixed-length style embedding using a trainable style encoder that is conditioned along with content features and speaker embeddings in an unsupervised manner [9]. Disentangling speech styles with a hierarchy of variational autoencoder (VAE) was introduced in [10, 11]. Estimating mutual information using Mutual-Information Neural Estimator (MINE) between the style and the content has been proposed in [12]. To improve the controllability in style modelling, fine-grained style transfer approaches were investigated in [13, 14, 15]. On the other hand, to facilitate semi-supervised approach, auxiliary style classification task was proposed to accurately capture style information from the reference utterances [16, 17]. The most recent studies in this direction also take into account the speaker identity [7, 18, 19] which may be hard to extend for TTS models where only a few seconds of target voices are available. In [20, 21], the authors combined speaker and style embeddings to build an all-round TTS system.

A universal TTS synthesis system is able to generate speech from text with speaker characteristics and speaking style similar to a given reference signal. The major challenge for universal TTS is speaker perturbation along with style transfer. The ultimate goal is to transplant prosody from arbitrary speakers, especially in the context of zero-shot learning where only a few seconds of data is available. Towards this aim, we employ a universal TTS (UTTS) framework which consists of four major components: content encoder, style encoder, speaker encoder and speech decoder. The content encoder generates a content embedding from the text. The style encoder represents the style factors into a style embedding, while the speaker encoder provides the speaker identity in the form of a speaker embedding. Finally, the speech decoder, conditioned on all the above embeddings, synthesizes the desired target speech.

When considering to generalize the models with multiple speakers and multiple styles using just the reconstruction loss, performance unfortunately deteriorates. During training, content information is leaked into the style embeddings ("content leakage") and speaker information into style embeddings ("style leakage"). Thus at inference, when the reference speech has different content from the input text, the decoder expects the content from the style vector ignoring some part of the content text. Moreover, speaker information could be expected from the style encoder leading to completely different speaker attribute.

To alleviate those issues, we suggest a novel **R**ényi **D**ivergence based **D**isentangled **R**epresentation (RDDR) algorithm. The minimization of Rényi divergence becomes feasi-

http://dx.doi.org/10.21437/Interspeech.2021-660

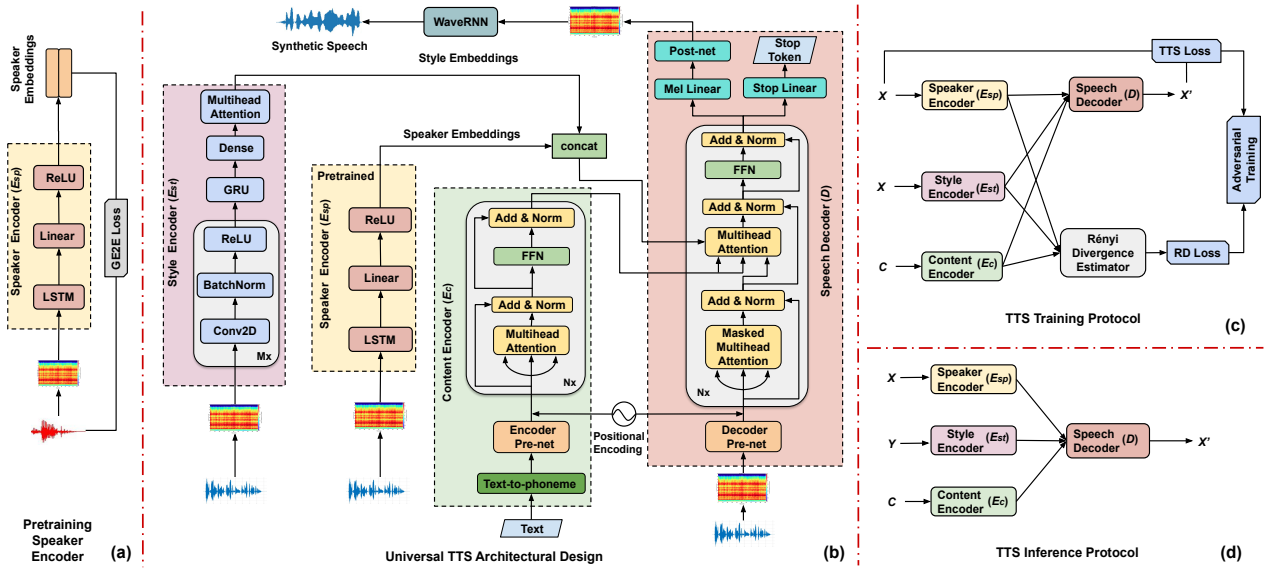

Figure 1: *System overview of our universal TTS framework. (a) Pre-training of the speaker encoder (E_{sp}) using generalized end-to-end loss (GE2E). (b) Universal TTS conditioned on speaker (E_{sp}) and style (E_{st}) encoders that can synthesized well controllable speech. TransformerTTS is employed as a backbone TTS infrastructure. (c) The proposed training protocol takes into account a novel adversarial RDDR approach combined with the minimization of the TTS reconstruction loss. A reference utterance is used to extract speaker and style factors, whereas during inference (d) the system may take as input any arbitrary speaker or style.*

ble via a variational representation formula which involves the cumulant generating function. We introduce two variations of this framework: Hellinger distance RDDR (H-RDDR) and sum of Rényi divergences RDDR (S-RDDR). Both variants are selected aiming to reduce the statistical variance of the adversarial component. Moreover, cumulants are preferred over expectations because they capture higher-order statistical information about the underlying distributions which often leads to more stable training [22]. Similar to mutual information minimization where a lower bound of the Kullback-Leibler divergence is utilized, the proposed RDDR algorithm estimates, via neural network approximations, a lower bound of the Rényi divergence between the joint distribution and the product of the marginals of two pairs: content-style and speaker-style and then, minimizes the estimated Rényi divergence in an adversarial manner. Rényi divergence minimization between those distributions pushes the various modalities to become independent. Our approach effectively disentangles content, style and speaker information that not only alleviates the leakage issues but also assists the decoder to be trained on the proper data leading to synthesizing high-quality speech. Our work involves independent training of a speaker-discriminative neural encoder to produce utterance level speaker embeddings using a state-of-the-art generalized end-to-end loss [23]. Hence the TTS system is capable of synthesizing speech from unseen speakers in a zero-shot manner. We train style encoder using a set of trainable vectors, which are linearly combined using style factors generated from the input reference speech. Style tokens are trainable parameters that are optimized together with the TTS network parameters. Finally, due to low computational complexity, TransformerTTS is employed for the content encoder and speech decoder [24]. Experimental results based on both objective and subjective evaluation confirm that the proposed method achieves better style similarity and perceptual speech quality than the baseline TTS system which is trained without a disentangled loss. Code and sound samples can be found in [1].

[1]https://dipjyoti92.github.io/Universal-TTS/

2. Universal TTS (UTTS)

An overview of the proposed universal TTS framework along with training and inference protocols, is shown in Figure 1.

2.1. Speaker Encoder

The Speaker Encoder (E_{sp}) deploys generalized end-to-end (GE2E) loss that is trained on thousands of speakers [23]. The log mel-spectrograms are extracted from speech utterances of arbitrary length. The feature vectors are then assembled in the form of a batch that contains S different speakers, and each speaker has U utterances. Each feature vector $\mathbf{f}_{ij}$ ($1 \leq i \leq S$ and $1 \leq j \leq U$) represents the features extracted from speaker i and utterance j, respectively. The features are then passed to an encoder architecture. The final embedding vector is L^2 normalized and they are calculated by averaging all window frames. The encoder consists of 3 LSTM layers of 768 cells followed by a projection to 256 dimensions as depicted in Figure 1(a). During training, embedding of all utterances for a particular speaker should be closer to the centroid of that particular speaker's embedding, and at the same time far from other speakers' centroid.

2.2. Style Encoder

The style encoder (E_{st}) is comprised of a convolutional stack, followed by a gated recurrent unit (GRU) similar to the GST-Tacotron paper [9]. Mel spectrograms, which are extracted from the reference speech, are passed to the stack of six 2D convolutional layers with kernel size 3×2 and 2×2 stride. The channel sizes of convolutional layers are (32, 32, 64, 64, 128, 128) followed by batch normalization and ReLU activation. The output from the last convolutional layer is summarized with a single-layer 128-unit unidirectional GRU. Style token layer is implemented with ten style token embeddings and a multi-head attention module [25]. As with speaker and content embeddings, the dimension of the style embeddings is 256. Finally, we apply $\tanh$ activation to GSTs before attention since it leads to greater token diversity. The overall style encoder is jointly trained with entire TTS model without using prosodic labels.

2.3. TTS Module

Given its performance and computational gains, we implemented TransformerTTS as our backbone TTS [24]. Here, the multi-head attention mechanism constructs the hidden states for the encoder and the decoder in parallel which improves training efficiency. In addition to this backbone TransformerTTS model, a style encoder E_{st} and a speaker encoder E_{sp} is introduced to construct a truly universal TTS, also depicted in Figure 1(b).

The TTS module converts textual, style and speaker information into acoustic features.The final module is a vocoder, WaveRNN [6] in our case, which generates speech waveform from the previously generated acoustic information. The first stage of the TTS module is the conversion from text to phonemes. The text-to-phoneme converter not only assists the model to train on vast majority of cases but also resolves cases where some letters can be pronounced differently under different context which can lead to major degradation in the performance when data are not sufficient enough. Given a set of speech and phoneme content pairs $(\mathbf{x}, \mathbf{c})$, the baseline UTTS minimizes the feature-domain reconstruction loss between the predicted output of the speech decoder D and the original speech,

$$\mathcal{L}_{tts} = \min_{E_{st},E_c,D} \parallel D(E_c(\mathbf{c}), E_{st}(\mathbf{x}), E_{sp}(\mathbf{x})) - \mathbf{x} \parallel_1 \quad (1)$$

where $||\cdot||_1$ is the L^1 norm. We are not optimizing E_{sp} weights due to the fact that it is already pre-trained on thousands of speakers. Therefore, the speaker embeddings should reflect a well-balanced speaker universe. TTS module is first pre-trained with LJSpeech which has a broad range of linguistic variability and then we freeze E_c for the remaining training process.

3. Proposed Disentangled Representation

Although, baseline UTTS tries to synthesize speech using content, style and speaker factors, training just on an L^1 reconstruction loss is not enough. The style embeddings still manage to carry non-style information, leading to content leakage and style leakage. To efficiently decouple all representations without explicit labels, we estimate and minimize the Rényi divergence (RD) between their embedding representation pairs $(E_c(\mathbf{c}), E_{st}(\mathbf{x}))$ and $(E_{sp}(\mathbf{x}), E_{st}(\mathbf{x}))$. The overall training and inference protocols are demonstrated in Figure 1(c) & (d).

3.1. Preliminaries

The standard approach to disentangle two modalities is through Mutual Information (MI) minimization since zero MI implies Independence. MI is the Kullback-Leibler (KL) divergence and it can be represented in the form of Donsker-Varadhan representation [26]. Given two random variable $\mathbf{X}$ and $\mathbf{Y}$, MI i.e., $\mathcal{I}(\mathbf{X}, \mathbf{Y})$ is equivalent to KL divergence between the joint distribution, $\mathbb{P}_{\mathbf{XY}}$ and the product of marginals, $\mathbb{P}_{\mathbf{X}} \otimes \mathbb{P}_{\mathbf{Y}}$. MINE [27] estimate a lower bound of MI:

$$\mathcal{I}(\mathbf{X}, \mathbf{Y}) \geq \mathcal{I}_\Theta(\mathbf{X}, \mathbf{Y}) = \sup_{\theta \in \Theta} \left\{ \mathbb{E}_{\mathbb{P}_{\mathbf{XY}}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_{\mathbf{X}} \otimes \mathbb{P}_{\mathbf{Y}}}[e^{T_\theta}]) \right\} \quad (2)$$

where T_θ is an NN-based parametrization of the space of all continuous and bounded functions. For the proposed UTTS, T_θ is parametrized by three fully connected layers, each layer followed by ReLU, with parameters $\theta \in \Theta$ while the optimization is performed through stochastic gradient descent.

We introduce this approach to disentangle content, style and speaker representations in UTTS, similar to [12]. We use two separate neural estimator T_θ and $T'_{\theta'}$ to approximate MI from the pairs $(E_c(\mathbf{c}), E_{st}(\mathbf{x}))$ and $(E_{st}(\mathbf{x}), E_{sp}(\mathbf{x}))$, respectively. By minimizing MI, we enforce the encoders to learn information that is independent of each other. Thus, the overall objective function is a min-max problem where we seek to maximize the lower-bound of MI w.r.t. T_θ and $T'_{\theta'}$ and minimize the MI and the reconstruction loss w.r.t. E_{st} and D.

$$\mathcal{L} = \min_{E_{st},D} \max_{T_\theta, T'_{\theta'}} \{ \parallel D(E_c(\mathbf{c}), E_{st}(\mathbf{x}), E_{sp}(\mathbf{x})) - \mathbf{x} \parallel_1 + \lambda \max(0, \mathcal{I}_\Theta(E_c(\mathbf{c}), E_{st}(\mathbf{x}))) + \lambda \max(0, \mathcal{I}_{\Theta'}(E_{st}(\mathbf{x}), E_{sp}(\mathbf{x})))\} \quad (3)$$

where λ is a hyper parameter, set to 0.1. We also bound from below the estimated MI to non-negative values.

3.2. Rényi Divergence based Disentangled Representation

Learning representative latent embeddings is a challenging problem. In order to decouple all the information factors properly, it is necessary to estimate the MI between the embedding pairs. However, it has been demonstrated that the statistical variance of the finite-sampling MI estimator can be exponentially high [28, 29] often resulting in inferior estimation performance. In this paper, we propose a different family of information-theoretic divergences aiming towards reduced estimator's variance. We present two alternatives based on the Rényi divergence family for disentangled speech representation learning. The proposed algorithm is presented in Figure 2.

Algorithm: Pseudo-code for proposed RDDR training

Input: Speech and text pairs $\langle \mathbf{x}_i, \mathbf{c}_i \rangle$.
Pre-training: Optimize E_c, D on LJSpeech using $\min_{E_c,E_{st},D} \sum_i \parallel D(E_c(\mathbf{c}_i), E_{st}(\mathbf{x}_i), E_{sp}(\mathbf{x}_i)) - \mathbf{x}_i \parallel_1$
$E_{sp} \leftarrow$ GE2E training
$E_{st}, T_\theta, T'_{\theta'} \leftarrow$ initialization with random weights
while $E_{st}, D, T_\theta, T'_{\theta'}$ *not converged* **do**
  Sample mini-batch from $\langle \mathbf{x}_i, \mathbf{c}_i \rangle$; i = {1, 2, . . . , b}
  $\{\mathbf{p}_i\} \leftarrow \{E_c(\mathbf{c}_i) | i = 1, 2, \ldots, b\}$
  $\{\mathbf{q}_i\} \leftarrow \{E_{st}(\mathbf{x}_i) | i = 1, 2, \ldots, b\}$
  $\{\mathbf{r}_i\} \leftarrow \{E_{sp}(\mathbf{x}_i) | i = 1, 2, \ldots, b\}$
  $\{\hat{\mathbf{p}}_i\}, \{\tilde{\mathbf{r}}_i\} \leftarrow$ random permutation of $\{\mathbf{p}_i\}$, $\{\mathbf{r}_i\}$
  $\mathcal{L}_{RD^1} = \sum_k [-\frac{1}{\beta_k} \log \frac{1}{b} \sum_{i=1}^b e^{-\beta_k T_\theta(\mathbf{p}_i, \mathbf{q}_i)} - \frac{1}{\gamma_k} \log \frac{1}{b} \sum_{i=1}^b e^{\gamma_k T_\theta(\hat{\mathbf{p}}_i, \mathbf{q}_i)}]$
  $\mathcal{L}_{RD^2} = \sum_k [-\frac{1}{\beta_k} \log \frac{1}{b} \sum_{i=1}^b e^{-\beta_k T'_{\theta'}(\mathbf{r}_i, \mathbf{q}_i)} - \frac{1}{\gamma_k} \log \frac{1}{b} \sum_{i=1}^b e^{\gamma_k T'_{\theta'}(\tilde{\mathbf{r}}_i, \mathbf{q}_i)}]$
  The overall objective function:
  $\mathcal{L} = \frac{1}{b} \sum_{i=1}^b \parallel D(\mathbf{p}_i, \mathbf{q}_i, \mathbf{r}_i) - \mathbf{x}_i \parallel_1 + \lambda \max(0, \mathcal{L}_{RD^1}) + \lambda \max(0, \mathcal{L}_{RD^2})$
  $D = D - \epsilon \nabla_D \mathcal{L}; \quad E_{st} = E_{st} - \epsilon \nabla_{E_{st}} \mathcal{L}$
  $T_\theta = T_\theta + \epsilon \nabla_D \mathcal{L}_{RD^1}; \quad T'_{\theta'} = T'_{\theta'} + \epsilon \nabla_D \mathcal{L}_{RD^2}$
end

Figure 2: *Training algorithm of RDDR.*

Given two random variable $\mathbf{X}$ and $\mathbf{Y}$, RDDR employs a DNN (i.e., T_θ) to approximate the maximum lower bound of the Rényi divergence (RD) variational formula which is defined via two cumulant generating functions (CGFs):

$$\mathcal{R}_{\beta,\gamma}(\mathbf{X}, \mathbf{Y}) \geq \sup_{\theta \in \Theta} -\frac{1}{\beta} \mathbb{E}_{\mathbb{P}_{\mathbf{XY}}}[e^{-\beta T_\theta}] - \frac{1}{\gamma} \log(\mathbb{E}_{\mathbb{P}_{\mathbf{X}} \otimes \mathbb{P}_{\mathbf{Y}}}[e^{\gamma T_\theta}]) \quad (4)$$

where hyper-parameters β and γ are two non-zero real numbers which control the learning dynamics. The use of CGFs allows for an inclusive characterization of the distributions' statistics, making it possible for T_θ to better enforce independence. This in turn leverages improved disentanglement representation. The proposed algorithm can be interpreted for several choices of its hyper-parameters. Thus, the optimization of the proposed loss function is equivalent to the minimization of a divergence for

a wide set of hyper-parameter values. For our experiments, we choose two variations. First, (β, γ) = (0.5, 0.5) which is equivalent to the minimization of Hellinger distance, we refer to this as Hellinger RDDR (H-RDDR). Internal numerical simulations conducted by our group have shown that the variance of the estimated Hellinger distance is significantly smaller than the variance of the estimated KL divergence. Second, we choose a combination of β = [0, 0.5, 1] and γ = [1, 0.5, 0] which is equivalent to the minimization of the sum of Rényi divergences. This variant is called sum of RDDR (S-RDDR). This particular choice for the hyper-parameters tries to optimize KL divergence, reverse KL and Hellinger distance simultaneously. Doing so, we anticipate enhanced independence between the speech factors. Similar to Equation (3), we can formulate the overall objective function in the form of adversarial training.

4. Results and Discussion

The speaker encoder training has been conducted on LibriSpeech, VoxCeleb1 and VoxCeleb2 datasets containing utterances from over 8k speakers [30]. TransformerTTS and WaveRNN models are trained using VCTK English corpus [31] from 109 different speakers. We initially train TransformerTTS model on LJSpeech database [32], which contains 13,100 audio clips from a single speaker, as a "warm-start" approach.

Table 1: *Objective evaluation tests. Lower scores indicate better performance.*

Methods	No Shuffle			Shuffle		
	RMSE-F0	MCD	WER(%)	RMSE-F0	MCD	WER(%)
UTTS	29.80	5.28	22.7	47.02	6.56	32.1
UTTS MINE	30.33	5.38	25.4	47.26	6.59	31.9
UTTS S-RDDR	**28.59**	5.35	21.6	**45.75**	**6.39**	28.7
UTTS H-RDDR	**28.59**	**5.26**	**18.3**	47.26	6.59	**26.6**

4.1. Objective Evaluation

In this section, we evaluate the performance of disentanglement strategies shown in Table 1. To assess the effectiveness of the proposed methods, we calculated three performance scores from 100 random samples. Mel-cepstral distortion (MCD) measures the spectral distance between the synthesized and reference mel-spectrum features. Root mean squared error (RMSE) evaluates the similarity in F0 modeling between reference and synthesized speech. Lastly, the content preservation criterion is evaluated by word error rate (WER). During inference, we evaluate the performance on two conditions: 'no shuffle' and 'shuffle'. No shuffle feeds same reference speech $\mathbf{x}_i$ into style and speaker encoders and its corresponding text $\mathbf{c}_i$ to predict the speech features $\mathbf{x}'$ with the decoder. Whereas, shuffle feeds speech $\mathbf{x}_i$ into speaker, $\mathbf{x}_j$ into style and $\mathbf{c}_k$ into the content encoder given ($i \neq j \neq k$). We observe that the proposed S-RDDR and H-RDDR algorithms outperform both baseline and MINE approach in terms of RMSE-F0 and MCD evaluation metrics with relative improvements. As expected, no shuffle scenarios perform better with respect to shuffled samples. Furthermore, one of the main objectives of RDDR algorithm is to improve the content leakage of the generated speech which we objectively measure using Google's open-source automatic speech recognizer (ASR) [33]. For the VCTK dataset, ASR achieves a WER of 14.6% on the held-out real data. Although, both RDDR variants perform better, the performance of H-RDDR is the best so far with WER of 18.3% and 26.6% for no shuffle and shuffle scenarios, respectively. We overall conclude that the disentanglement module during training assists the TTS to achieve more accurate rendering of prosodic patterns as well as synthesizing proper speech content to its corresponding text without any significant leakage issues.

Table 2: *Average cosine-similarity evaluation.*

Methods	Baseline	MINE	S-RDDR	H-RDDR
No Shuffle	0.828	0.840	0.836	0.839
Shuffle	0.734	0.732	0.737	0.739

Next, we employ cosine-similarity as a speaker similarity measure between the generated and reference speaker's speech. As shown in Table 2, across different systems, cosine-similarity does not vary much. It is attributed to the fact that speaker embeddings are pre-trained and donot jointly train with the TTS module. Therefore, different TTS modules perform equally better and speaker identities are much closer to reference speaker.

Table 3: *MOS scores (95% confidence interval) of audio quality and speaking style similarity for different TTS modules.*

Methods	MOS-Speech-Quality	MOS-Style-Similarity
UTTS	3.01 ± 0.05	2.98 ± 0.08
UTTS MINE	3.11 ± 0.06	2.92 ± 0.08
UTTS S-RDDR	**3.62** ± 0.05	**3.41** ± 0.07
UTTS H-RDDR	3.51 ± 0.06	3.36 ± 0.07

4.2. Subjective Evaluation

We conduct listening tests to evaluate different TTS modules, and the choice of RDDR approach as depicted in Table 3. Twenty native and non-native English listeners participated in our listening tests. We conducted two separate mean opinion score (MOS) listening tests and subjects were asked to rate the synthesized speech on a scale of five-point (1:Bad, 2:Poor, 3:Fair, 4:Good, 5:Excellent). MOS-Speech-Quality (MSQ) assesses the perceptual speech quality whereas MOS-Style-Similarity (MSS) evaluates the speaking style expressiveness w.r.t. the reference style. MSQ scores indicate that compared to UTTS, the proposed S-RDDR and H-RDDR UTTSs are superior in quality with 20.3% and 16.6% relative improvement respectively. We also found that our proposed TTS mimic better style characteristics than baseline and shows significant relative improvement of 14.4% and 12.7% for S-RDDR and H-RDDR respectively. Results indicate that disentanglement helps the system to properly learn all the information factors relating to content, speaker and style. Therefore, it enhances the speech decoder's ability to synthesize high-quality speech and also preserves the style of the reference speech better. We did not conduct listening tests for speaker similarity as objective results clearly indicate equal performance for all TTS modules.

5. Conclusions

We proposed a novel disentangled representation by exploiting cumulant generating functions in speech synthesis. Our system approximates and then minimizes the Rényi divergence between content-style and style-speaker pairs, and it is jointly trained with TTS reconstruction loss in an adversarial manner. Subjective and objective evaluation revealed that the proposed approach outperforms both the baseline and MINE algorithm and is able to eliminate the issues of content and style leakage, resulting in a truly universal TTS system. The main advantage of universal TTS is its high controllability, since it improves multi-speaker multi-style training along with better generalization ability by allowing reliable transfer to speaker and style information. Furthermore, speaker and style conditioning can be computed using few seconds of data in an unsupervised manner.

6. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[3] S. Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *International Conference on Machine Learning*. PMLR, 2017, pp. 195–204.

[4] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," *arXiv preprint arXiv:2006.04558*, 2020.

[5] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[6] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.

[7] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen *et al.*, "Hierarchical generative modeling for controllable speech synthesis," *arXiv preprint arXiv:1810.07217*, 2018.

[8] 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 *Proc. international conference on machine learning*. PMLR, 2018, pp. 4693–4702.

[9] 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.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[11] A. Tjandra, R. Pang, Y. Zhang, and S. Karita, "Unsupervised learning of disentangled speech content and style representation," *arXiv preprint arXiv:2010.12973*, 2020.

[12] T. Y. Hu, A. Shrivastava, O. Tuzel, and C. Dhir, "Unsupervised style and content separation by minimizing mutual information for speech synthesis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3267–3271.

[13] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5911–5915.

[14] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," *arXiv preprint arXiv:1907.02479*, 2019.

[15] T. Daxin and L. Tan, "Fine-grained style modelling and transfer in text-to-speech synthesis via content-style disentanglement," *arXiv preprint arXiv:2011.03943*, 2020.

[16] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 623–627.

[17] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[18] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *arXiv preprint arXiv:1806.04558*, 2018.

[19] S. Ma, D. Mcduff, and Y. Song, "A generative adversarial network for style modeling in a text-to-speech system," in *International Conference on Learning Representations*, vol. 2, 2019.

[20] C. M. Chien, J. H. Lin, C. Y. Huang, P. C. Hsu, and H. Y. Lee, "Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech," *arXiv preprint arXiv:2103.04088*, 2021.

[21] D. Tan, H. Huang, G. Zhang, and T. Lee, "CUHK-EE voice cloning system for ICASSP 2021 M2VoC challenge," *arXiv preprint arXiv:2103.04699*, 2021.

[22] Y. Pantazis, D. Paul, M. Fasoulakis, Y. Stylianou, and M. Katsoulakis, "Cumulant GAN," *arXiv preprint arXiv:2006.06625*, 2020.

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

[24] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[26] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time. IV," *Communications on Pure and Applied Mathematics*, vol. 36, no. 2, pp. 183–212, 1983.

[27] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *International Conference on Machine Learning*. PMLR, 2018, pp. 531–540.

[28] J. Song and S. Ermon, "Understanding the limitations of variational mutual information estimators," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1x62TNtDS

[29] D. McAllester and K. Stratos, "Formal limitations on the measurement of mutual information," in *Proceedings of Machine Learning Research*. PMLR, 26–28 Aug 2020, pp. 875–884. [Online]. Available: http://proceedings.mlr.press/v108/mcallester20a.html

[30] D. Paul, Y. Pantazis, and Y. Stylianou, "Speaker conditional WaveRNN: Towards universal neural vocoder for unseen speaker and recording conditions," *Proc. Interspeech 2020*, pp. 235–239, 2020.

[31] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2016.

[32] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[33] "Google's speech-to-text," https://cloud.google.com/speech-to-text.