

MULTI-SPEAKER EXPRESSIVE SPEECH SYNTHESIS VIA MULTIPLE FACTORS DECOUPLING

Xinfa Zhu, Yi Lei, Kun Song, Yongmao Zhang, Tao Li, Lei Xie*

Audio, Speech and Language Processing Group (ASLP@NPU)
School of Computer Science, Northwestern Polytechnical University, Xi'an, China

ABSTRACT

This paper aims to synthesize the target speaker's speech with desired speaking style and emotion by transferring the style and emotion from reference speech recorded by other speakers. We address this challenging problem with a two-stage framework composed of a text-to-style-and-emotion (Text2SE) module and a style-and-emotion-to-wave (SE2Wave) module, bridging by neural bottleneck (BN) features. To further solve the multi-factor (speaker timbre, speaking style and emotion) decoupling problem, we adopt the multi-label binary vector (MBV) and mutual information (MI) minimization to respectively discretize the extracted embeddings and disentangle these highly entangled factors in both Text2SE and SE2Wave modules. Moreover, we introduce a semi-supervised training strategy to leverage data from multiple speakers, including emotion-labeled data, style-labeled data, and unlabeled data. To better transfer the fine-grained expression from references to the target speaker in non-parallel transfer, we introduce a reference-candidate pool and propose an attention-based reference selection approach. Extensive experiments demonstrate the good design of our model.

Index Terms— Expressive speech synthesis, multiple factors decoupling, two-stage, style transfer, emotion transfer

1. INTRODUCTION

In recent years, neural text-to-speech (TTS) synthesis has made rapid progress regarding quality and naturalness [1, 2, 3, 4]. With the wide applications of TTS, there have been increasing demands for robust expressive speech synthesis systems to provide more human-like speech in diverse scenarios. In previous works of expressive speech synthesis, speech expressiveness usually refers to specific speaking styles or emotional expressions associated with speech [5, 6, 7, 8].

A straightforward approach [9, 10, 11, 12] to synthesize expressive speech for a specific speaker is to train a TTS system with his/her expressive training speech. However, it can not be generalized to target speakers without expressive training data, which is hard to obtain for each target speaker. Therefore, transferring emotion or style from a source speaker to target speakers is a feasible strategy, where the source speaker has expressive speech while the target speaker only has neutral speech. The key factor for emotion or style transfer is to decouple the speaker timbre and expressive aspects from speech [13, 14, 8, 15]. Speaker timbre essentially reveals the physiological characteristics of individual's vocal tract, while emotion and speaking style are more behavioral. However, it is not a trivial task as these aspects are highly entangled in the speech signal. Some works [6, 16, 17] try to disentangle speaker timbre and style or emotion in the latent space to conduct style or emotion transfer. However, decoupling approaches in latent space usually need to carefully select an appropriate reference signal during inference.

These reference-based style transfer methods always face a trade-off between expressiveness and speaker similarity, which leads to a vast performance gap between synthetic and natural human speech.

To solve this problem, some articles adopt the neural network bottleneck (BN) features or Phonetic PosteriorGrams (PPGs) as intermediate representations for decoupling. BN features are a set of activation of nodes over time from a neural network bottleneck layer, while PPGs are obtained by stacking a sequence of phonetic posterior probabilistic vectors from the neural network. BN features and PPGs, usually obtained from a well-trained acoustic model in an automatic speech recognition system, are believed to be linguistically rich [18, 19], speaker-independent [20], noise-robust [21], and contain stylistic information such as duration and accent [22]. Leveraging the intermediate representations, the style transfer TTS problem can be simplified to a two-stage process, where the first stage mainly manages the style learning from the source speaker while the second stage aims at the target speaker timbre modeling. Through such a two-stage framework, style or emotion transfer can be conducted without a reference signal during inference. Referee [8] is a representative work in this direction that adopts PPGs as the intermediate representations connecting the two-stage models for cross-speaker style transfer.

The above style transfer approaches usually have an unclear definition of style and emotion and sometimes consider emotion as a type of speaking style. Whereas, this indiscriminate treatment of style and emotion restricts them to be extended to diverse scenarios requiring both multiple emotions and styles, which is common in real applications. Speaking style is a general distinctive style of speech in different usage scenarios, such as news reading, storytelling, poetry recitation, and conversation. Even for storytelling, telling different stories (such as fairy tales and novels) may use different speaking styles. By contrast, emotion mainly reflects the mood state of the speaker, related to attitudes and intentions, conveyed differently in each utterance, such as happy, angry, sad, etc. Moreover, different emotions can be expected in different places in an audio stream with a global speaking style (such as storytelling).

In this paper, we focus on both speaking style and emotion transfer in multi-speaker expressive speech synthesis. The challenges for building such a multi-factor system are threefold. First, explicitly decoupling multiple factors – speaking style, emotion, and speaker timbre – is more difficult as style and emotion patterns are both reflected in speech prosody and thus highly entangled. Second, it is also difficult to obtain expressive data labeled with both emotion and speaking style. Third, the reference-based model mentioned above has a mismatch problem in practical non-parallel transfer scenarios, i.e. the novel text content during inference is different from the reference signal selected in the training data, leading to performance degradation, which is severe in multi-speaker expressive speech synthesis [14] and may become more critical in our multi-factor case.

* Corresponding author. This work was supported by National Key R&D Program of China, under Grant No. 2020AAA0108600.

Our proposed approach leverages the advances of the two-stage framework with a text-to-style-and-emotion module (*Text2SE*) and a style-and-emotion-to-wave (*SE2Wave*) module. The former predicts linguistic, style, and emotional information embedded in the neural bottleneck (BN) feature, while the latter takes the BN feature as input and produces the target speaker’s voice with both stylistic and emotional factors.

Based on the two-stage framework, this paper proposes the following designs. To address the multi-factor decoupling problem, we adopt the multi-label binary vector (MBV) [23] and mutual information (MI) minimization [24] to respectively discretize the extracted embeddings and decouple the style, emotion, and speaker factors in the design of both *Text2SE* and *SE2Wave* modules. As for the data sparsity problem, we introduce a semi-supervised training strategy to leverage expressive data from multiple speakers, including emotion-labeled data, style-labeled data, and unlabeled data. To eliminate the mismatch problem in non-parallel transfer scenarios, we introduce a reference-candidate pool and propose an attention-based reference selection approach, which reserves the fine-grained prosody from the reference signal and avoids the difficulty of manual reference selection. Extensive experiments demonstrate the good design of our model. We suggest the readers listen to our online demos¹.

2. PROPOSED APPROACH

The proposed approach consists of a *Text2SE* module and a *SE2Wave* module, as shown in Figure 1. The *Text2SE* module is to predict BN features, pitch, and energy, which are speaker-independent intermediate representations with style and emotion. The *SE2Wave* module aims at waveform generation of the target speaker in the desired emotion and style. Note that emotional information is superposed in the procedure of wave generation as the supplement of the fine-grained emotion variations for natural expression delivery. As detailed in Section 2.4, the whole system will go through a training stage and a fine-tuning stage, where different embedding extractors are used in the two stages respectively to ensure good performance.

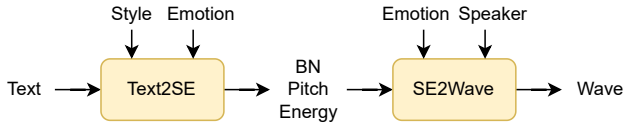


Fig. 1. The architecture of proposed approach.

2.1. The *Text2SE* module

As illustrated in Figure 2, the *Text2SE* module is shaped with a backbone model and two embedding extractors for style and emotion respectively. The backbone model is mainly composed of a phoneme encoder, a variance adaptor, and a BN decoder. The goal of this module is to produce the speaker-independent BN features conditioning on the style and emotion embeddings.

During training, the style/emotion embedding extractor contains a style/emotion encoder, a Multi-label Binary Vector (MBV) [23], and a classifier. The style/emotion encoder takes the mel-spectrogram as input and then uses an MBV to discretize the output for compression. As a bottleneck layer, MBV with Gumbel-Softmax can reduce the difficulty of multi-factor decoupling and stabilize the adversarial training. To train the embedding extractors, we add classification constraints to the extracted embeddings. Besides, we adopt the variational contrastive log-ratio upper bound (vCLUB) [24] to measure the mutual information between the extracted style and emotion embeddings for sufficiently decoupling style and emotion by mutual information (MI) minimization.

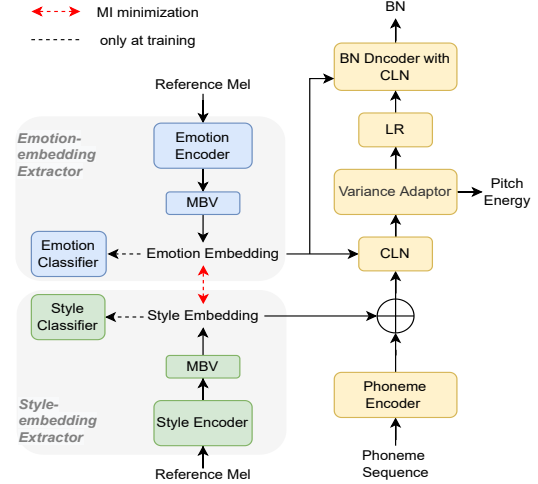


Fig. 2. The *Text2SE* module architecture.

The extracted emotion embedding is fed into the backbone model through Conditional LayerNorm (CLN), and the style embedding is first concatenated with the phoneme encoder output and then goes through the CLN. The variance adaptor predicts the pitch and energy normalized in the utterance level for eliminating the speaker-related attributes. Finally, the BN decoder produces the speaker-independent BN with style and emotional information. The backbone network follows the structure used in FastSpeech2 [2].

The training objective of the *Text2SE* module \mathcal{L}_{t2se} is

$$\mathcal{L}_{t2se} = \mathcal{L}_{BN} + 0.1 \cdot \mathcal{L}_{prosody} + 0.1 \cdot \mathcal{L}_{MI} + \mathcal{L}_{emo} + \mathcal{L}_{sty}, \quad (1)$$

where \mathcal{L}_{BN} is the reconstruction loss of BN, $\mathcal{L}_{prosody}$ is the loss for predicting pitch and energy, \mathcal{L}_{MI} is the MI loss between emotion embedding and style embedding, \mathcal{L}_{emo} and \mathcal{L}_{sty} are the classification loss of emotion and style embedding.

2.2. The *SE2Wave* module

Likewise, the *SE2Wave* architecture is composed of a backbone model and two embedding extractors for speaker and emotion respectively. The backbone is based on VITS [4], consisting of a BN encoder, Flow, posterior encoder, HiFi-GAN decoder, and discriminator, as shown in Figure 4. The BN encoder, conditioned on the extracted emotion embedding, takes the BN feature, pitch, and energy as input to provide the prior distribution of the speaker-independent representations. The speaker embedding from the speaker lookup table with MBV is treated as the conditional constraint of the Flow model. Similar to the *Text2SE* module, for decoupling the speaker and emotion in the *SE2Wave* module, we also add classification constraints to the emotion embedding and use vCLUB to minimize mutual information between the emotion embedding and the stop gradient speaker embedding.

We denote \mathcal{L}_{emo} as the emotion classification loss and $\mathcal{L}_{MI'}$ as the MI loss in the *SE2Wave* module. The L1 loss is used as the reconstruction loss \mathcal{L}_{rec} to minimize the mel-spectrogram of the ground truth and predicted waveform. The adversarial training loss and feature map loss in VITS [4] are also adopted in the model for improving the performance of wave generation.

The training objectives of the *SE2Wave* module are as follows:

$$\mathcal{L}_{se2w}^G = \mathcal{L}_{kl} + 45 \cdot \mathcal{L}_{rec} + 0.1 \cdot \mathcal{L}_{MI'} + \mathcal{L}_{emo} + \mathcal{L}_{adv}(G) + \mathcal{L}_{fm}(G) \quad (2)$$

$$\mathcal{L}_{se2w}^D = \mathcal{L}_{adv}(D), \quad (3)$$

where \mathcal{L}_G , \mathcal{L}_D , and \mathcal{L}_{kl} are the generative loss, discriminator loss, and KL divergence of the hidden distribution.

¹Demo: <https://zxf-icpc.github.io/multi-factor-decoupling/>

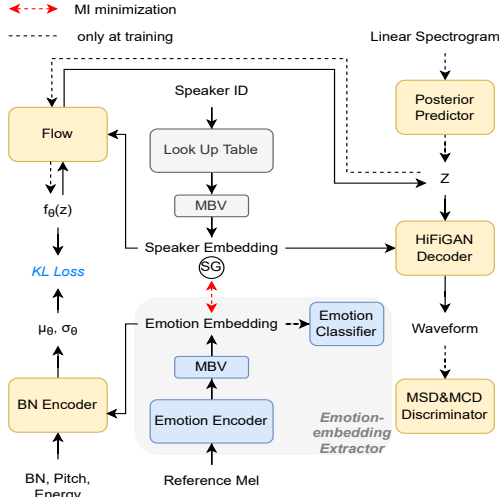


Fig. 3. The SE2Wave module architecture.

2.3. Semi-supervised training

Suppose we have a variety of multi-speaker expressive data at hand, including emotion-labeled data, style-labeled data, and unlabeled data. To better leverage all of the data, we introduce a simple semi-supervised training strategy to train the emotion/style embedding extractors for multi-label classification. Specifically, for the emotion-labeled data, only \mathcal{L}_{emo} is calculated and \mathcal{L}_{sty} is not considered in Eq. (1); the model softly determines what style it should be. Likewise for the style-labeled data. As for the unlabeled data, neither \mathcal{L}_{emo} nor \mathcal{L}_{sty} is calculated in Eq. (1) and Eq. (2), and the model will softly determine what style or emotion it contains.

2.4. Attention based reference selection

At training time, the style/emotion embedding is extracted from the target mel-spectrogram, which is parallel with the text content. While during inference, the reference signal is different from the novel text (i.e. non-parallel). This results in a mismatch problem of the extracted embedding between the training and inference stages, leading to degraded performance according to previous studies [14]. To relieve the performance degradation brought by the mismatch, we introduce an extra stage to fine-tune the Text2SE and SE2Wave modules. Particularly in the fine-tuning stage, we introduce a novel embedding extractor to replace the original emotion/style extractor used in the previous training stage. The new extractor aims to alleviate the aforementioned mismatch problem and select the appropriate reference in a soft way. Thus the new embeddings are used as conditions for the Text2SE and SE2Wave modules.

As illustrated in Figure 4, the embedding extractor in fine-tuning procedure employs scaled dot-product attention [25] to calculate the embedding output as the conditional constraints for the Text2SE and SE2Wave modules. We introduce an *embedding-candidate pool* providing the candidates as the attention *key* and *value*, retrieved from the given style/emotion ID. Specifically, the embedding-candidate pool consists of N embeddings for each category (e.g., emotion-happy or style-fairy-tales) extracted by the style/emotion encoder and MBV in the previous training stage. Unlabeled data is treated as a special category. The attention *query* is provided by a GST-layer [26] from the input hidden representations which are the phoneme encoder outputs in the Text2SE module and BN, pitch, and energy inputs in the SE2Wave module respectively. In this way, the attention mechanism selects the embedding with the most significant attention weight as the appropriate reference embedding for the

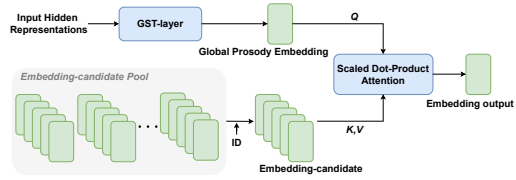


Fig. 4. The embedding extractor in fine-tuning procedure.

2.5. Pipeline

The pipeline of the proposed approach contains training, fine-tuning, and inference phases. During training, we train the two-stage model using the objectives mentioned above and update the emotion-/style-embedding extractors E_{emo}^t and E_{sty}^t by the extracted embeddings with the constraint of MI loss and the semi-surprised classification loss. During fine-tuning, the embedding extractors E_{emo}^f and E_{sty}^f are updated by the objectives of the acoustic model. In inference, the proposed two-stage system adopts the same embedding extractor as the fine-tuning stage. Both the Text2SE and SE2Wave modules automatically select the appropriate embedding according to the textual input, style, and emotion.

3. EXPERIMENTS

3.1. Experimental Setup

Three internal Mandarin corpora are involved in the experiments: 1) dataset **M30S3** contains 30 speakers, and its total duration is 18.5 hours, in which each speaker has 1 to 3 styles including *poetry recitation*, *story telling - fairy tales* and *story telling - novels*; 2) dataset **M3E6** contains three speakers, and its total duration is 21.1 hours, in which each speaker has six emotions of *anger*, *fear*, *happiness*, *sadness*, *surprise* and *neutral*; and 3) dataset **M30U** has 30 speakers with neither style tags nor emotion tags, and its total duration is 18.2 hours. For all recordings, the sample rate is converted to 24 kHz. The BN features are extracted with 12.5ms hop-size and 256-dimension from a robust TDNN-F model trained with 30k hours of Mandarin speech data. To validate the performance of our proposed approach, we implement the following systems:

- **MR-Tacotron:** Multi-reference structure follows [14] to disentangle and control specific styles based on the FastSpeech 2 architecture for a fair comparison. A HiFi-GAN [3] vocoder is adopted to transform the predicted mel-spectrogram into speech waveform.
- **Referee:** A cross-speaker style transfer framework follows Referee [8] with additional emotion transfer.
- **Proposed:** the two-stage framework proposed in this paper to decouple and recompose the multiple factors in speech.

In our implementation of the proposed approach, the emotion and style encoders follow the structure of mel-style encoder proposed by Meta-StyleSpeech [27]. All classifiers have the same structure that consists of 3 fully connected layers. The Text2SE backbone and mutual information estimator follow the settings of FastSpeech 2 [2], and VQMVC [28] respectively. The SE2Wave backbone follows the settings of VITS [4], and the BN encoder consists of 6 FFT blocks. During fine-tuning, we set $N = 100$ for the embedding-candidate pool in this paper.

3.2. Subjective evaluation

We conduct mean opinion score (MOS) experiments to evaluate speech naturalness, emotion similarity, speaker similarity, and style similarity. Specifically, given 20 reserved transcripts for each style, we generate samples respectively for each emotion category, resulting in 360 listening samples ($20 \times 3 \times 6$). We invite 20 native

Table 1. Results of subjective evaluation with 95% confidence interval and objective evaluation.

Model	Naturalness	Emotion Similarity	Speaker Similarity	Style Similarity	CER (%)	Cosine Similarity
Proposed	4.03 ± 0.08	3.89 ± 0.10	3.93 ± 0.07	3.81 ± 0.11	6.7	0.856
MR-Tacotron [14]	3.83 ± 0.09	3.38 ± 0.12	3.62 ± 0.10	3.77 ± 0.10	6.0	0.804
Referee [8]	3.07 ± 0.11	2.88 ± 0.12	3.37 ± 0.08	3.43 ± 0.11	7.6	0.846

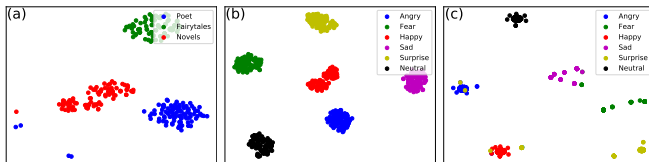
Mandarin Chinese speakers to participate in the listening tests. In each test session (naturalness/emotion/speaker/style), participants are asked to rate how similar the synthetic and the reference speech is in the specific assessment metric while ignoring other aspects.

As shown in Table 1, the proposed approach significantly outperforms Referee and MR-Tacotron in terms of speech naturalness, emotion similarity, and speaker similarity. The high emotion and speaker similarity scores demonstrate that the proposed approach can decouple the emotion and speaker identity effectively. For the style similarity, the proposed approach achieves much better performance than Referee and is slightly better than MR-Tacotron. The results of the emotion, style and speaker similarity indicate that the proposed method can effectively decouple multiple factors from speech. Besides, the proposed method achieves the best audio naturalness, indicating its flexibility and stability in generating specific emotional speech of target speakers.

3.3. Objective evaluation

Robustness. We measure the character error rate (CER) of the synthesized samples by the pre-trained WeNet [29] to assess the robustness of the models. Moreover, we use the pre-trained ECAPA-TDNN [30] to extract the x-vector and calculate the cosine similarity to further verify the speaker similarity. As shown in Table 1, the proposed model achieves the highest cosine similarity. Interestingly, the Referee achieves a similar cosine similarity to the proposed method. However, subjective tests show a massive gap between the proposed method and the Referee in speaker similarity. We speculate that the poor audio quality of the Referee affects the listeners’ judgment. We observe that the ASR model does not recognize Poet well, where CER is generally high. Considering speech generated from the proposed method is expressive of the multiple factors, it’s reasonable to get a slightly higher CER than MR-Tacotron.

Effectiveness of semi-supervised training. To verify the effectiveness of semi-supervised training, we visualize the emotion and style embeddings through t-SNE [31]. One hundred utterances reserved per emotion or style are adopted for the test. As shown in Figure 5(a) and 5(b), the style and emotion embeddings are well clustered, proving the effectiveness of semi-supervised training. Moreover, Figure 5(c) demonstrates that emotion embeddings in the SE2Wave model cannot form clusters by categories. We conjecture that emotion embeddings in SE2Wave mainly focus on and supplement the fine-grained emotional variance since emotion in Text2SE represents the global category.

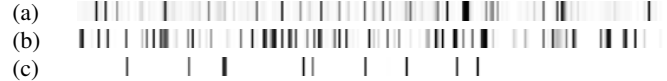
**Fig. 5.** T-SNE results of style embedding and emotion embedding. (a) style embedding in Text2SE, (b) emotion embedding in Text2SE and (c) emotion embedding in SE2Wave.

Emotion and Style Embedding. We extract embeddings of all training utterances and then calculate the standard deviation of all embeddings for each dimension. The lower standard deviation

Table 2. Results of Ablation study with 95% confidence interval.

Model	Naturalness	Emotion Similarity	Speaker Similarity	Style Similarity
Proposed	4.03 ± 0.08	3.89 ± 0.10	3.93 ± 0.07	3.81 ± 0.11
- MBV	3.47 ± 0.10	3.68 ± 0.08	3.33 ± 0.10	3.60 ± 0.14
- MI	3.74 ± 0.10	3.88 ± 0.08	3.21 ± 0.11	3.71 ± 0.14
- FT	3.88 ± 0.08	3.83 ± 0.09	3.63 ± 0.09	3.76 ± 0.13
- EE	3.85 ± 0.08	3.49 ± 0.09	3.90 ± 0.08	3.78 ± 0.14
- BN	2.63 ± 0.12	2.47 ± 0.13	2.20 ± 0.11	2.38 ± 0.15

means the dimension changes less among embeddings of all utterances, which is less valuable, in other words. The embeddings are intuitively presented in Figure 6, where the darker vertical bar means a higher standard deviation. We can observe that some dimensions are purely white, meaning they are not used at all. Moreover, Figure 6(c) shows obviously fewer dimensions are used in the SE2Wave emotion embedding since it is just a supplement to provide fine-grained emotional variance. These observations show that MBV can effectively discretize information in the embeddings. More importantly, Figure 6(a) and (b) are mutually exclusive, indicating that style and emotion are well decoupled.

**Fig. 6.** Results of the standard deviation of each dimension of the extracted embedding. (a) style embedding in Text2SE, (b) emotion embedding in Text2SE and (c) emotion embedding in SE2Wave.

3.4. Ablation study

We conduct an ablation study to validate the components of our proposed method by removing certain modules and jointly training two modules, as shown in Table 2. The results show that removing the Multi-label Binary Vector (-MBV) module leads to a decline in perceptive evaluations, indicating that MBV improves system stability. Removing the Mutual information loss (-MI) and the emotion extractor in SE2Wave (-EE) lead to a sharp decline in speaker similarity and emotion similarity, respectively, highlighting the effectiveness of MI loss in decoupling multiple factors and the importance of fine-grained emotional variance composition in the SE2Wave stage. The results (-FT) also show that the fine-tuning process can effectively improve overall performance. Furthermore, the jointly trained model (-BN) fails to disentangle multiple factors, resulting in synthetic audio with low naturalness, expressiveness, and speaker similarity.

4. CONCLUSIONS

This paper aims to synthesize speech with desired style and emotion for a target speaker by transferring the style and emotion from reference speech recorded by other speakers. We approach this challenging problem with a two-stage framework composed of a text-to-style-and-emotion (Text2SE) module and a style-and-emotion-to-wave (SE2Wave) module, while neural bottleneck features are served as the intermediate representation. Importantly, based on this framework, we have proposed several contributions, including multi-factor decomposition, semi-supervised training to better leverage data, and attention-based reference selection. Extensive experiments demonstrate the good design of our model.

5. REFERENCES

- [1] Xu Tan, Tao Qin, Frank K. Soong, and Tie-Yan Liu, “A survey on neural speech synthesis,” 2021, vol. abs/2106.15561.
- [2] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” in *Proc. ICLR*. 2021, OpenReview.net.
- [3] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Proc. NIPS*, 2020.
- [4] Jaehyeon Kim, Jungil Kong, and Juhee Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *Proc. ICML*. 2021, pp. 5530–5540, PMLR.
- [5] Yinghao Aaron Li, Cong Han, and Nima Mesgarani, “Styletts: A style-based generative model for natural and diverse text-to-speech synthesis,” 2022, vol. abs/2205.15439.
- [6] Alexander Sorin, Slava Shechtman, and Ron Hoory, “Principal style components: Expressive style control and cross-speaker transfer in neural TTS,” in *Proc. Interspeech*. 2020, pp. 3411–3415, ISCA.
- [7] Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie, “Mse-motts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 853–864, 2022.
- [8] Songxiang Liu, Shan Yang, Dan Su, and Dong Yu, “Referee: Towards reference-free cross-speaker style transfer with low-quality data for expressive speech synthesis,” in *Proc. ICASSP*. 2022, pp. 6307–6311, IEEE.
- [9] Younggun Lee, Azam Rabiee, and Soo-Young Lee, “Emotional end-to-end neural speech synthesizer,” 2017, vol. abs/1711.05447.
- [10] Runnan Li, Zhiyong Wu, Yuchen Huang, Jia Jia, Helen Meng, and Lianhong Cai, “Emphatic speech generation with conditioned input layer and bidirectional LSTMs for expressive speech synthesis,” in *Proc. ICASSP*. 2018, pp. 5129–5133, IEEE.
- [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*. 2019, pp. 6945–6949, IEEE.
- [12] Tao Li, Shan Yang, Liumeng Xue, and Lei Xie, “Controllable emotion transfer for end-to-end speech synthesis,” in *Proc. ISCSLP*. 2021, pp. 1–5, IEEE.
- [13] Jie Wang, Jingbei Li, Xintao Zhao, Zhiyong Wu, Shiyin Kang, and Helen Meng, “Adversarially learning disentangled speech representations for robust multi-factor voice conversion,” in *Proc. Interspeech*. 2021, pp. 846–850, ISCA.
- [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,” 2019, vol. abs/1904.02373.
- [15] Yi Lei, Shan Yang, Xinfu Zhu, Lei Xie, and Dan Su, “Cross-speaker emotion transfer through information perturbation in emotional speech synthesis,” 2022, vol. 29, pp. 1948–1952.
- [16] Shuang Ma, Daniel McDuff, and Yale Song, “Neural TTS stylization with adversarial and collaborative games,” in *Proc. ICLR*. 2019, OpenReview.net.
- [17] Tao Li, Xinsheng Wang, Qicong Xie, Zhichao Wang, and Lei Xie, “Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 1448–1460, 2022.
- [18] Damien Ronssin and Milos Cernak, “AC-VC: non-parallel low latency phonetic posteriorgrams based voice conversion,” in *Proc. ASRU*. 2021, pp. 710–716, IEEE.
- [19] Haohan Guo, Zhiping Zhou, Fanbo Meng, and Kai Liu, “Improving adversarial waveform generation based singing voice conversion with harmonic signals,” in *Proc. ICASSP*. 2022, pp. 6657–6661, IEEE.
- [20] Yen-Hao Chen, Da-Yi Wu, Tsung-Han Wu, and Hung-yi Lee, “Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization,” in *Proc. ICASSP*. 2021, pp. 5954–5958, IEEE.
- [21] Zhonghao Li, Benlai Tang, Xiang Yin, Yuan Wan, Ling Xu, Chen Shen, and Zejun Ma, “Ppg-based singing voice conversion with adversarial representation learning,” in *Proc. ICASSP*. 2021, pp. 7073–7077, IEEE.
- [22] Houjun Huang, Xu Xiang, Yexin Yang, Rao Ma, and Yanmin Qian, “Aispeech-sjtu accent identification system for the accented english speech recognition challenge,” in *Proc. ICASSP*. 2021, pp. 6254–6258, IEEE.
- [23] Andy T. Liu, Po-chun Hsu, and Hung-yi Lee, “Unsupervised end-to-end learning of discrete linguistic units for voice conversion,” in *Proc. Interspeech*. 2019, pp. 1108–1112, ISCA.
- [24] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin, “CLUB: A contrastive log-ratio upper bound of mutual information,” in *Proc. ICML*. 2020, pp. 1779–1788, PMLR.
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Proc. NIPS*, 2017, pp. 5998–6008.
- [26] Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. 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 *Proc. ICML*. 2018, pp. 5167–5176, PMLR.
- [27] Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang, “Meta-stylespeech : Multi-speaker adaptive text-to-speech generation,” in *Proc. ICML*. 2021, pp. 7748–7759, PMLR.
- [28] Disong Wang, Liquan Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng, “VQMIVC: vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion,” in *Proc. Interspeech*. 2021, pp. 1344–1348, ISCA.
- [29] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei, “Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit,” in *Proc. Interspeech*. 2021, pp. 4054–4058, ISCA.
- [30] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, “ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in *Proc. Interspeech*. 2020, pp. 3830–3834, ISCA.
- [31] Laurens van der Maaten and Geoffrey E. Hinton, “Visualizing data using t-sne,” 2008, vol. 9, pp. 2579–2605.