

# Multi-reference Tacotron by Intercross Training for Style Disentangling, Transfer and Control in Speech Synthesis

Yanyao Bian, Changbin Chen, Yongguo Kang, Zhenglin Pan

Baidu Speech Department  
Baidu Inc. Baidu Technology Park, Beijing, 100193, China

{biyananyao, chenchangbin, kangyongguo, panzhenglin}@baidu.com

## Abstract

Speech style control and transfer techniques aim to enrich the diversity and expressiveness of synthesized speech. Existing approaches model all speech styles into one representation, lacking the ability to control a specific speech feature independently. To address this issue, we introduce a novel multi-reference structure to Tacotron and propose intercross training approach, which together ensure that each sub-encoder of the multi-reference encoder independently disentangles and controls a specific style. Experimental results show that our model is able to control and transfer desired speech styles individually. **Index Terms:** expressive speech synthesis, style disentangling, multi-style control

## 1. Introduction

Recent years neural text-to-speech (TTS) has been developing rapidly and has become an indispensable part of our daily life. End-to-end-TTS (E2E) models, which generate speech without handcraft linguistic features, are one of the most attractive TTS techniques and have achieved great progresses in voice quality [1, 2, 3]. However, it remains a challenge for E2E models to control speech styles, such as speaker, emotion, prosody, etc., which are essential for expressive and diverse voice generation.

Currently there are mainly two ways to control and transfer speech styles. The supervised way, which takes attributes' id as an additional model input, has been proved effective in multi-speaker TTS [4]. However, this approach can hardly deal with more complex styles like emotion and prosody, because there are no objective measurements to annotate these styles. The unsupervised way, which combines E2E models and reference encoders into Autoencoder (AE) [5] or Variational Autoencoder (VAE) [6, 7, 8], gets closer results. Theoretically, the unsupervised E2E models can model any complex styles in a continuous latent space, so that one can control and transfer style by manipulating the latent variables.

However, most of the previous researches model all speech styles into one style representation, which contains too much interfering information to be robust and interpretable. When conducting style transfer, one has to transfer all styles whether desired or not, which may not fit the contexts thus hurts generalization. When conducting style control, one can hardly confirm the relationship between the styles and the coefficients of each dimension of the style representations. Therefore, a natural question then arises: can we control different styles independently in an interpretable way?

To address the above issues, (1) we introduced a multi-reference encoder to GST-Tacotron [9], a state-of-the-art E2E model for style disentangling, control and transfer, to model multiple styles simultaneously. (2) We introduced intercross

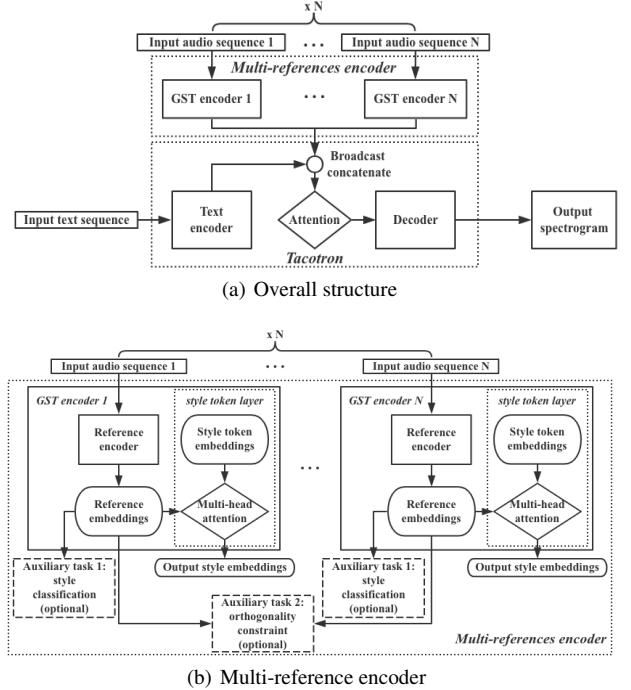


Figure 1: Structure of the proposed model.

training (IT) to extract and separate different classes of speech styles. Several experiments are conducted to show the proposed methods' feasibility.

The rest of the paper is organized as follows. Section 2 introduces multi-reference structure, intercross training, auxiliary tasks and inference procedures. Section 3 presents experimental results. Section 4 gives a conclusion of this paper.

## 2. Model and metrics

### 2.1. Model structure

Our model is based on GST-Tacotron. Figure 1(a) shows its overall structure, which mainly consists of two parts: Tacotron and multi-reference encoder. Figure 1(b) shows the detailed structure of multi-reference encoders, which consists of  $N$  GST sub-encoders, abbreviated as sub-encoders. The Tacotron and GST-encoder in this paper have no differences from those in [9]. We use Griffin-Lim [10] for fast experiment and use WaveNet [11] to generate demos for better audio fidelity.

For each sub-encoder, the reference encoder takes a reference audio (or acoustic parameters) as input and outputs the ref-

*erence embeddings*. Then a multi-head attention is conducted between reference embeddings and the *style token embeddings* to generate the *style embeddings*. Finally the style embeddings of each sub-encoder are broadcast and concatenated to the outputs of text encoder for standard Tacotron decoding. In addition, we add two optional auxiliary tasks in multi-reference encoder for training our models, which will be discussed later.

## 2.2. Definitions

For better explaining how to control multiple styles, we firstly introduce the definitions of styles in this paper.

The *style class* is the abstract description of styles which we want to disentangle and control. In our experiments, we evaluate 3 style classes: speaker, emotion, prosody. However, the proposed method is not constrained on these style classes.

The *style instance* refers to the instances of style classes. In this paper, speaker class is instantiated as 300 different speakers; emotion class is instantiated as happy, sad, angry, fear, confuse, surprise and neutral; prosody class is instantiated as news, story, radio, poetry and call-center.

In this paper, we use  $X_{i_1 i_2 \dots i_N, j}$  to represent an audio (or acoustic parameters) with  $N$  style classes to discuss, each of which instantiated as  $i_1, i_2, \dots, i_N$ , and with text content  $j$ . Furthermore, we introduce a wildcard “\*” to represent “any”. For example,  $X_{*i_n*, j}$  represents an audio with style instance  $i_n$ , text content  $j$  and a number of arbitrary other style instances.

Each sub-encoder in the multi-reference encoder is designed to model only one style class. Formally, we want the following equation (1) to be maintained in each sub-encoder.

$$Q_{\phi_n}(S_{i_n} | X_{*i_n*, j}) = Q_{\phi_n}(S_{i_n} | X_{*i_n*, k}) \quad (1)$$

for  $n = 1, 2, \dots, N$

Where  $S_{i_n}$  represents style instance  $i_n$ ;  $\phi_n$  represents the parameters of n-th sub-encoder.

This equation is saying that the n-th sub-encoder provides the same posterior distribution given audios with the same style instance of the n-th style class, regardless of other style classes and text contents. For example, in a two-reference model, we can have 1-st sub-encoder focus on modeling speaker style class and generating speaker embeddings, while utilizing the 2-nd sub-encoder to deal with prosody style class. Consequently, we can control different style classes independently by manipulating their own style embeddings.

## 2.3. Intercross training

Consider the simplest case  $N = 1$ . If we pick a pair of audios  $\{X_{i_1, j}, X_{i_1, k}\}$  with the same style instance  $i_1$  and text contents  $j \neq k$ , simultaneously minimizing following reconstruction loss functions to zero guarantees the equation (1) holds:

$$\begin{aligned} \mathcal{L}_{ORG}(X_{i_1, j}, X_{i_1, k}; \theta, \phi_1) &= \\ &- E_{Q_{\phi_1}(S_{i_1} | X_{i_1, j})} [\log P_{\theta}(X_{i_1, j} | S_{i_1}, c_j)] \quad (2) \\ &- E_{Q_{\phi_1}(S_{i_1} | X_{i_1, k})} [\log P_{\theta}(X_{i_1, k} | S_{i_1}, c_k)] \end{aligned}$$

$$\begin{aligned} \mathcal{L}_{IT}(X_{i_1, j}, X_{i_1, k}; \theta, \phi_1) &= \\ &- E_{Q_{\phi_1}(S_{i_1} | X_{i_1, k})} [\log P_{\theta}(X_{i_1, j} | S_{i_1}, c_j)] \quad (3) \\ &- E_{Q_{\phi_1}(S_{i_1} | X_{i_1, j})} [\log P_{\theta}(X_{i_1, k} | S_{i_1}, c_k)] \end{aligned}$$

Where  $\theta$  represents the parameters of Tacotron. Equation (2) is the original object whose reference and target are the same, and equation (3) are called intercross object as we exchange the  $Q_{\phi}(S | X)$  when doing reconstruction. After fully optimizing the two objects, we see  $\mathcal{L}_{ORG} = \mathcal{L}_{IT}$  ideally. The posterior distributions of different audios  $P_{\theta}(X | S, c)$  are different, hence the exchanged part must be equal, then equation (1) holds.

Notice that equation (2) can be included in equation (3) by removing the constraint  $j \neq k$ . Furthermore, with the assumption that a fully training procedure would have enumerated all reference-target combinations, we could optimize the two parts of equation (3) separately in different training steps as an approximation. Therefore, the final objective function of  $N = 1$  intercross training comes to the following equation (4).

$$\mathcal{L}_{IT} = -E_{Q_{\phi_1}(S_{i_1} | X_{i_1, *})} [\log P_{\theta}(X_{i_1, j} | S_{i_1}, c_j)] \quad (4)$$

To minimize equation (4), we randomly pick two audios with replacement from a group of audios  $\mathcal{G} = \{X_{i_1, *}\}$ , where audios in  $\mathcal{G}$  share the same style instance  $i_1$ . Then we use one of them as reference and the other as target to train the model.

For  $N > 1$  cases, equation (4) can be generalized to equation (5) according to the multi-reference structure, which takes  $N$  styles from  $N$  reference inputs to construct the target.

$$\begin{aligned} \mathcal{L}_{IT} &= -E_{\prod_{n=1}^N Q_{\phi_n}(S_{i_n} | X_{*i_n*, *})} \\ &[\log P_{\theta}(X_{i_1 i_2 \dots i_N, j} | S_{i_1 i_2 \dots i_N}, c_j)] \quad (5) \end{aligned}$$

To minimize equation (5), we firstly randomly pick a target  $T = X_{i_1 i_2 \dots i_N, j}$ . Then we randomly pick  $N$  references  $R = \{X_{i_1*, *}, X_{*i_2*, *}, \dots, X_{*i_N*, *}\}$  with replacement, whose n-th element  $R_n = X_{*i_n*, *}$  shares the same style instance  $i_n$  with  $T$ . Finally we construct the references-target pair ( $R, T$ ) as a sample to train the model.

We have realized that the concept of intercross training has been mentioned before in [12]. However, they did not discuss its general form for multi-style disentangling in AE structures.

## 2.4. Auxiliary tasks

In our experiments, our model did not converge well because it was confused by the meanings of multi-reference encoder style embeddings. To address this issue, we introduce a style classification task  $\mathcal{L}_{classification}$  over the reference embeddings. Intuitively, the style classification task ensures that the desired style information has been already modeled in the reference embeddings, and it will be easier for intercross training to squeeze out other redundancies.

Inspired by [13], we further introduce an orthogonality constraint shown in equation (6) to strengthen the independence of the style embeddings of each sub-encoder.

$$\mathcal{L}_{orthogonality} = \sum_{ij} \|H_i^T H_j\|_F^2 \quad (6)$$

Where  $\|\cdot\|_F^2$  is the squared Frobenius norm and  $H_i, H_j$  means the reference embeddings of the i-th and j-th sub-encoder, respectively. The final object comes to equation (7):

$$\mathcal{L} = \mathcal{L}_{IT} + \beta \mathcal{L}_{classification} + \gamma \mathcal{L}_{orthogonality} \quad (7)$$

In our experiments, we set  $\beta$  to 1 and  $\gamma$  to 0.02 without exploring. Notice that the auxiliary tasks are optional and only performed when  $N > 1$ .

## 2.5. Inference

Before experiments, we would like to introduce the inference procedures of the proposed model. We present an example of 2-references model with sub-encoder 1 to control speaker and sub-encoder 2 to control prosody. The inference procedures are shown in Figure 2.

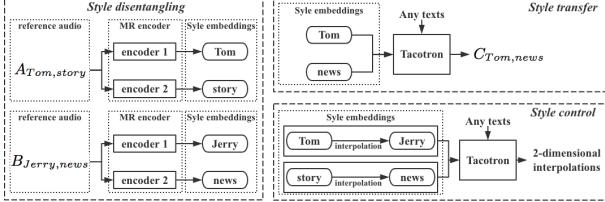


Figure 2: *Inference procedure of style disentangling, transfer and control using the proposed model.*

**Style disentangling.** The fundamental of this paper described in equation (1) and achieved by IT. Given a audios, each sub-encoder can model a specific style.

**Style transfer.** Different disentangled style embeddings can be freely combined with each other to generate new voices.

**Style control.** Styles can be controlled through linear interpolation between style embeddings of style instances in each encoder, formulated as equation (8).

$$SE_{to} = SE_{from} + \alpha \cdot (SE_{to} - SE_{from}) \quad (8)$$

Where  $\alpha \in [0, 1]$ ;  $SE$  refers to style embeddings.

**Random sampling.** The proposed model is able to generate speech with random styles following equation (9).

$$SE_{random} = \sum_{k=1}^K softmax(a_k) \cdot STE_k \quad (9)$$

Where  $STE$  means the style token embeddings in a sub-encoder, shaped  $K \times D$ ;  $a_k \sim \mathcal{U}(0, 1)$ . Speech generated in this way could be with any style instances of the style class controlled by current sub-encoder.

## 3. Experiments

In this section, we firstly build a single-reference ( $N = 1$ ) model to evaluate the performance of intercross training (IT). Then we build a multi-reference ( $N = 2$ ) model to show the extendibility for controlling multiple styles.

In the following experiments, all speech data are recorded by BAIDU Speech Department and the sampling rate is 16 kHz. The model takes 80-dimensional mel-spectrogram with 50ms frame length and 12.5ms frame shift as reference input and output. Meanwhile, it is conditioned on Mandarin phonemes and tones. The model outputs 5 frames per decoding step.

As the visualization in speech synthesis is often not enough for objective perception, we strongly encourage readers to listen to the audio samples provided on our demo page<sup>1</sup>.

### 3.1. Single-reference

In single-reference experiments, we mainly analyze the results of speaker style. The results of prosody style are presented in

<sup>1</sup><https://ttsdemos.github.io/paper3.html>

multi-reference experiments, and the results of emotion style can be found in our demo page.

We train models over an 110 hours corpus recorded by 300 different non-professional Chinese speakers including 178 females and 122 males, each of which recorded about 500 utterances. We use the original GST model (ORG) as comparison.

#### 3.1.1. Style disentangling

We randomly chose 20 speakers including 10 females and 10 males and visualize the style embeddings of their about 10000 recordings using t-SNE [14] on Tensorboard shown in Figure 3.

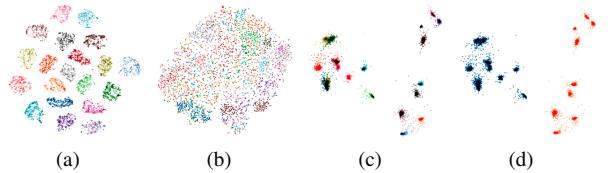


Figure 3: *Style embeddings visualization. Pictures are colored by speakers if not specified. (a) IT-SE by t-SNE; (b) ORG-SE by t-SNE; (c) IT-SE by PCA. (d) IT-SE colored by genders.*

Compared ORG-SE in Figure 3(b), we can clearly distinguish all 20 speakers with IT-SE in Figure 3(a). That is to say, unlike ORG-SE that contains lots of redundancies, IT-SE concentrate on speaker style class, which is consistent with equation (1). Furthermore, we also visualize the same IT-SE by PCA, which is a linear dimension reduction method, in Figure 3(c) and 3(d). The results indicate that the IT-SE of different speakers and genders are linearly separable.

#### 3.1.2. Style transfer

We conduct non-parallel style transfer by randomly choosing 4 various audios from a speaker as reference audios to synthesize a fixed unseen text. Figure 4 shows the results.

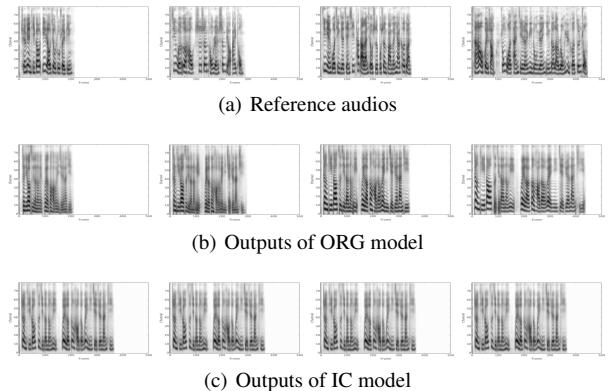


Figure 4: *Non-parallel style transfer conditioned on same text and different reference audios from same speaker.*

We see that the output audios of the ORG model in Figure 4(b) are with the lengths coinciding with reference audios instead of texts, and both naturalness and speaker similarity with reference audios are limited. In contrast, the output audios of the IT model in Figure 4(c) are with similar lengths coinciding with texts, which sound natural and similar to the reference.

The results indicate that disentangled style embeddings ensure the success of non-parallel style transfer.

### 3.1.3. Style control

We randomly chose two audios from a female speaker A and a male speaker B as references and implement linear interpolation introduced in Section 2.5 over style embeddings from A to B. The results are shown in Figure 5.

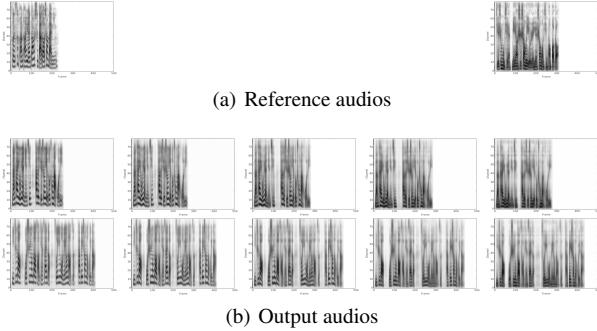


Figure 5: Style control by linear interpolation. (b) Left to right: referencing style embeddings interpolating from A to B with  $\alpha$  set to 0.0, 0.25, 0.5, 0.75, 1.0. Top to bottom: Conditioning on Chinese sentences with 16 tokens, 24 tokens

We see a smooth conversion from female speaker A to male speaker B in speaker related features such as F0 and prosody simultaneously. The results indicate that the style latent space is continuous and can be interpolated freely. Furthermore, style transfer could be seen as special cases ( $\alpha = 0$  or 1) of style control so it was omitted in multi-reference experiments.

### 3.1.4. Few-shot/one-shot speaker conversion

Few-shot/one-shot speaker conversion is an important application in multi-speakers TTS. In our models, it is a special case of style transfer, where the reference audios are out of the training set. As we have achieved linear interpolation, we deduce that if the styles of the extra speakers can be represented as linear combinations of the styles of the existing speakers, then one-shot conversion can be done correctly. And we achieved about 20% accept rate by one-shot. Then we randomly picked 20 utterances from the failed speaker to fine-tuning the model as few-shot. We found that the model performs best when fine-tuning both multi-reference encoder and decoder while keeping the text encoder fixed. We achieved 100% accept rate by few-shot. Experiments details are presented in our demo page.

## 3.2. Multi-references

In these experiments, we built 2-reference models to control two style classes: speaker and prosody. Models were trained on a 30 hours corpus with 27 speakers and 5 prosodies. Partial details of the corpus is presented in Table 1.

### 3.2.1. Style disentangling

We randomly select 100 utterances from each prosody of each speaker, and visualize their style embeddings by t-SNE. Figure 6 shows the results.

We see the embeddings from speaker encoder in Figure 6(a) are clustered only by speaker, while the embeddings from

Table 1: Partial details of multi-style (speaker-prosody) data corpus. Empty means no data. M = male, F = female.

speaker	news	story	radio	poetry	call-center
M2				402	
M5	423	440	453	388	
F4					421
F17	1987		467	376	464
Total	3613	7101	3812	2893	3013

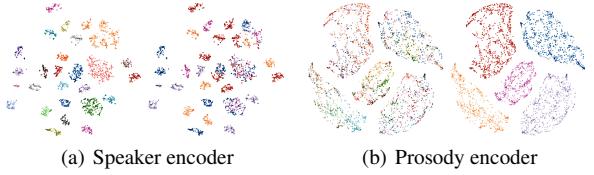


Figure 6: Visualize style embeddings of two sub-encoders. In each sub-figure, the left picture is colored by speakers and the right picture is colored by prosodies.

prosody encoder in Figure 6(b) are clustered only by prosody. The results prove that the speaker and prosody styles are disentangled well into designated encoders, so that we can control these styles independently.

### 3.2.2. Style control

We conducted style control over both parallel and non-parallel speakers. Here “non-parallel” means the speakers have no common target prosody styles, vice versa. Figure 7 shows the non-parallel results over M2-poetry and F4-callcenter. The text contents are fixed for better distinguishing the styles.

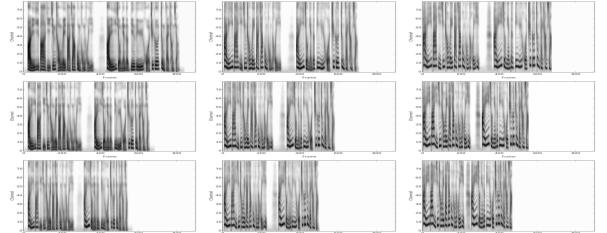


Figure 7: multi-style control over non-parallel data. M2 to F4 from left to right, poetry to callcenter from top to bottom.  $\alpha$  of both speaker and prosody is set to 0.0, 0.5, 1.0.

We see gradual and independent changes of F0 from male to female and audio lengths from poetry to callcenter. The results indicate that different style latent spaces are independent from each other, while each style latent space is continuous, which ensures the success of multi-dimensional linear interpolation.

Furthermore, random sampling and few-shot/one-shot conversion mentioned in Section 2.5 and 3.1.4 can also be applied in any sub-encoders of the multi-reference models to synthesize more expressive and diverse voices.

## 4. Conclusions

In this paper, we introduced multi-reference encoder to Tacotron and proposed intercross training technique. We prove that our model could disentangle different speech style classes into corresponding reference encoders and could control them independently. We showed that the proposed model could accomplish style disentangling, control, transfer, and other style related tasks in a flexible, interpretable and robust manner.

## 5. References

- [1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.
- [2] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, “Deep voice 3: Scaling text-to-speech with convolutional sequence learning,” *arXiv preprint arXiv:1710.07654*, 2017.
- [3] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, “Close to human quality tts with transformer,” *arXiv preprint arXiv:1809.08895*, 2018.
- [4] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, “Deep voice 2: Multi-speaker neural text-to-speech,” in *Advances in neural information processing systems*, 2017, pp. 2962–2970.
- [5] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” *arXiv preprint arXiv:1803.09047*, 2018.
- [6] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, “Expressive speech synthesis via modeling expressions with variational autoencoder,” *arXiv preprint arXiv:1804.02135*, 2018.
- [7] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, “Learning latent representations for style control and transfer in end-to-end speech synthesis,” *arXiv preprint arXiv:1812.04342*, 2018.
- [8] 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.
- [9] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” *arXiv preprint arXiv:1803.09017*, 2018.
- [10] D. Griffin and J. Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.
- [11] A. v. d. 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,” *arXiv preprint arXiv:1609.03499*, 2016.
- [12] D. Sun, T. Ren, C. Li, H. Su, and J. Zhu, “Learning to write stylized chinese characters by reading a handful of examples,” *arXiv preprint arXiv:1712.06424*, 2017.
- [13] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, “Domain separation networks,” in *Advances in Neural Information Processing Systems*, 2016, pp. 343–351.
- [14] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.