Enriching Source Style Transfer in Recognition-Synthesis based Non-Parallel Voice Conversion

Zhichao Wang[1], Xinyong Zhou[1], Fengyu Yang[1], Tao Li[1], Hongqiang Du[1], Lei Xie[1],*
Wendong Gan[2], Haitao Chen[2], Hai Li[2]

[1]Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science,
Northwestern Polytechnical University, Xi'an, China
[2]iQIYI Inc, China

`zcwang_aslp@mail.nwpu.edu.cn,{xyzhou,fyyang,taoli,hqdu}@nwpu-aslp.org,`
`lxie@nwpu.edu.cn,{wendonggan,haitaochen,haili}@qiyi.com`

Abstract

Current voice conversion (VC) methods can successfully convert timbre of the audio. As modeling source audio's prosody effectively is a challenging task, there are still limitations of transferring source style to the converted speech. This study proposes a source style transfer method based on recognition-synthesis framework. Previously in speech generation task, prosody can be modeled explicitly with prosodic features or implicitly with a latent prosody extractor. In this paper, taking advantages of both, we model the prosody in a hybrid manner, which effectively combines explicit and implicit methods in a proposed prosody module. Specifically, prosodic features are used to explicit model prosody, while VAE and reference encoder are used to implicitly model prosody, which take Mel spectrum and bottleneck feature as input respectively. Furthermore, adversarial training is introduced to remove speaker-related information from the VAE outputs, avoiding leaking source speaker information while transferring style. Finally, we use a modified self-attention based encoder to extract sentential context from bottleneck features, which also implicitly aggregates the prosodic aspects of source speech from the layered representations. Experiments show that our approach is superior to the baseline and a competitive system in terms of style transfer; meanwhile, the speech quality and speaker similarity are well maintained.

Index Terms: voice conversion, style transfer, hybrid modeling

1. Introduction

Voice conversion (VC) aims to modify speech from a source speaker to sound like that of a target speaker without changing the linguistic content. Various voice conversion approaches have been proposed before, including Gaussian mixture models (GMM) [1, 2], frequency warping [3, 4] and exemplar methods [5, 6]. These methods usually require parallel data from the source and target speakers, which is expensive to collect. Therefore, many works [7, 8] propose to use non-parallel data for voice conversion. Benefiting from the deep learning's strong ability of extracting rich features and non-linear regression, non-parallel approaches based on deep learning have become the mainstream. Among them, phonetic posteriorgram (PPG) [8–10] based approaches have received much attention. The framework generally consists of a *recognition* stage and a *synthesis* stage, where PPG or neural network bottleneck features (BN) serve as the 'information bridge'. For the recogni-

tion stage, an ASR acoustic model is used to extract speaker-independent PPG or BN features, which represents the linguistic content. The synthesis stage is a conversion model, which maps the extracted features by ASR to the acoustic features of the target speaker. Although sequence-to-sequence framework has been recently proposed to integrate the recognition and synthesis stages into a unified encoder-decoder neural architecture that can be trained in an end-to-end manner [11–13], the classical recognition-synthesis framework is still the widely-adopted non-parallel voice conversion framework in real applications because it is *flexible* – the two stages can be separately trained and *robust* – the ASR acoustic model can be trained using plenty of data.

Despite recent progress, most voice conversion methods focus on the transformation of timbre to the target and the preservation of linguistic content of the source. However, the prosody of the source is also desired to be correctly transferred to the target. The prosody information at least includes emotion, pitch, duration, and loudness, which are important in some scenarios for voice conversion with expressive source speech, such as dubbing, live broadcasting and data enhancement. Therefore, transferring prosody of source speech to converted speech is an interesting task.

In speech generation tasks, including both voice conversion and text-to-speech, modeling prosody has two directions in general: prosody can be modeled explicitly with prosodic features, or implicitly with a latent prosody extractor. The explicit modeling methods extract prosodic features [14–16] from speech to represent prosody information. As prosodic features are intuitive, it is easy to understand the effect on prosody, and the prosody can be controlled independently via varying features. However, prosody is related to many factors, it is difficult to describe prosody through handcrafted acoustic features perfectly. For implicit modeling, the prosody extractor can be GST [17, 18], reference encoder [19–21] and VAE [22], which extract comprehensive prosody embedding from reference speech. However, the extracted prosody embedding is difficult to explain and control the different aspects of prosody. Additionally, the embedding may contain other prosody unrelated information, e.g. timbre, referred as timbre leakage. To solve this problem, gradient reversal layer (GRL) [10, 20, 23–26] is introduced to remove speaker-related information from prosody.

In this paper, to obtain a non-parallel VC model with good performance and robustness in source style transfer, we choose a recognition-synthesis framework. To solve the prosody modeling problems caused by using explicit or implicit modeling
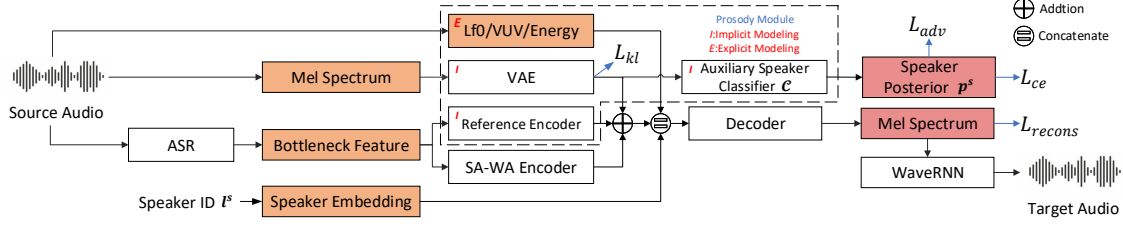
---

*Corresponding author.

Figure 1: *Overview of the components of the proposed conversion model.*

alone, we design a hybrid prosody module to model prosody, which effectively combines explicit modeling and implicit modeling methods. The prosodic features are used to explicit model prosody, including logarithmic domain fundamental frequency (lf0) and the short-term average amplitude (energy), which model pitch and energy. According to the characteristic that mel spectrum is speaker-dependent and BN is speaker-independent but prosodic-dependent, we use VAE and reference encoder [19–22] to implicitly model prosody, which take mel spectrum and BN as input respectively. Adversarial training strategy [13, 24] is also applied to wipe out the speaker-related information in the VAE output. To the best of our knowledge, we are the first to use BN to extract prosody representation. Additionally, we use the weighted aggregation self-attention encoder (SA-WA) [27] to extract sentential context from BN, which also implicitly aggregates the prosodic aspects of source speech from the layered representations. Experimental results show that with the proposed methods enriching source style transfer both explicitly and implicitly, our framework is clearly superior than the baseline systems.

The rest of the paper is organized as follows: Section 2 introduces our proposed methods. Section 3 presents the experiments and the comparison of our proposed approach against the baseline and ablation systems in terms of both subjective and objective measures. Section 5 concludes this paper.

2. Proposed approach

Our proposed source style transfer framework is shown in Fig. 1. It is based on an encoder-decoder architecture consisting of three components: a prosody module, an SA-WA encoder and a decoder. The prosody module learns to extract speaker-independent prosody representations. The SA-WA encoder exploits sentential context, which is used to excavate the sentential context from different levels. The decoder takes prosody, content, and speaker embedding as input and the output is mel spectrum. Finally WaveRNN [28] is adopted to reconstruct waveform from mel spectrum.

2.1. Prosody module

As shown in Fig. 1, to extract more comprehensive prosody representations, we designed a prosody module using explicit and implicit hybrid modeling methods.

Explicit modeling: prosody includes duration, pitch, energy, and so on. To finely characterize and control the prosody, we extract acoustic features from the source audio as conditions for the model. Frame level's VC ensures the duration information's consistency. For pitch and energy, we extract lf0 and the short-term average amplitude (energy), respectively. We also extract the vuv, which indicates the voicing of the current frame. In implementation, we use World [29] to extract lf0. To reduce

the differences between the extracted features caused by different speakers, we use the min-max normalization to normalize lf0 and energy to [0,1], which is useful to prevent performance degradation caused by the unseen style during the inference phase.

Implicit modeling: we use implicit modeling method consisting of two parts: VAE followed by an auxiliary speaker classifier and reference encoder. Inspired by [10, 18, 20, 21], we adopt VAE to extract the prosodic representation $\mathbf{z}$ from source audio's mel spectrum. We use the same VAE model's architecture and hyperparameters as [22]. As $\mathbf{z}$ produced by mel spectrum is related to the speaker, which affects the timbre of the converted speech, we adopt an speaker classifier $\mathcal{C}$ to utilize adversarial training [13, 24] to remove speaker-related information. The process of adversarial training can be described as:

$$p^s = \mathcal{C}(\mathbf{z}) \tag{1}$$

$$L_{ce} = \text{CE}(p^s, l^s) \tag{2}$$

$$L_{adv} = ||p^s - e||_2^2 \tag{3}$$

where l^s is the one-hot speaker label and p^s is the prediction of speaker classifier. The auxiliary speaker classifier is trained with cross-entropy loss L_{ce}. To make the $\mathbf{z}$ indistinguishable to the speaker classifier, we minimize the L_{adv} to force the speaker prediction p^s to obey a uniform distribution $e = [1/S, \ldots, 1/S]$, S is the number of speakers. Although adversarial training successfully removes speaker-related information in $\mathbf{z}$, part of prosody information is also inevitably removed. From [8, 24], we can learn that BN is a speaker-independent representation, which contains prosody information as well. We consider using a reference encoder [19] to extract speaker-independent prosodic representation directly from BN to enhance prosody information further. The outputs of VAE and reference encoder are concatenated to form the prosody embedding of our implicit modeling method.

2.2. SA-WA encoder

As shown in Fig. 2, we adopt an self-attention weighted aggregation (SA-WA) encoder [27] to learn a comprehensive sentence representation, which is aggregated by the prosodic-related sentential context information from different self-attention layers. There are three parts in the SA-WA encoder: encoder prenet, self-attention module and weighted aggregation. The architecture of encoder prenet is the same as the settings in Tacotron [30]. The self-attention module contains N self-attention blocks and each block consists of two sub-networks: muti-head attention and a feed forward network. Layer normalization and residual connection are also used in each sub-network. For the nth self-attention block, the relations among previous block f^{n-1}, first sub-network's output m^n and second sub-network f^n can be described as:

$$m^n = \text{LN}(\text{MultiHead}(f^{n-1}) + f^{n-1}), \qquad (4)$$

$$f^n = \text{LN}(\text{FFN}(m^n) + m^n) \qquad (5)$$

where MultiHead($\cdot$), FFN($\cdot$) and LN($\cdot$) are multi-head attention, feed forward network and layer normalization respectively. The weighted aggregation module consists of two sub-networks, similar to the self-attention block. Weighted aggregation can be formulated as follows:

$$g^n = \text{MeanPool}(\text{Conv1d}(f^n)), \qquad (6)$$

$$c = \text{LN}(\text{Multihead}(g^1, ..., g^N) + g^N), \qquad (7)$$

$$g = \text{LN}(\text{FFN}(c) + c), \qquad (8)$$

where Conv1d means 1d-convolution, MeanPool represents mean pooling, the modified c offers a more precise control of aggregation for each g^n, and g is the final output of encoder.
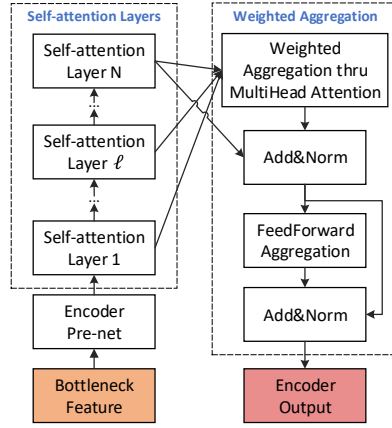

Figure 2: *The network architecture of the SA-WA encoder*

2.3. Loss function

The whole conversion model is jointly optimized. We regard the auxiliary speaker classifier as discriminator D, and the remaining parts as generator G, as shown in Fig. 1. The overall loss function can be described as follows:

$$\begin{cases} Loss_G = L_{recons} + \beta L_{adv} + \gamma L_{kl} \\ \\ Loss_D = L_{ce} \end{cases} \qquad (9)$$

where L_{recons} denotes the mel spectrum reconstruction loss, β is the weight for the adversarial loss, and γ is the variable weight for Kullback-Leibler (KL) loss to prevent VAE collapse. G and D are trained alternately. In general implementation, the discriminator D is updated several times for each update of the generator G. It is crucial to ensure that the classifier has a certain classification ability in the adversarial training process. To improve the model's generalization ability and the speaker similarity of converted speech, the whole network is first pre-trained on a multi-speaker dataset, and then is adapted with data from the target speaker. Note that only the decoder is optimized during the adaptation process.

3. Experiments

3.1. Dataset and Experimental Setup

A multi-speaker corpus containing 42 hours of speech from 59 speakers is used to pre-train the conversion model. A female speaker database *db1* [31] is utilized to fine-tune the conversion model. The langue is Mandarin Chinese. For conversion test, 240 utterances are used to evaluate, which contains different styles, including reading, emotion, news, novels, live broadcast, and dubbing. We use 80-dim mel spectrum as features computed in 50ms frame length and 12.5ms frame shift. Our SI-ASR acoustic system is a robust TDNN-F model trained with 30k hours of Mandarin speech data. We use the 256-dim bottleneck features as the linguistic representation, which is extracted from the last fully-connected layer before softmax. WaveRNN [28] is adopted to reconstruct waveform from mel spectrum. The vocoder is trained using db1 speech data.

To validate the performance of our proposed method, we implement the following systems:

- **CS:** we implement a comparison system in [21].
- **BL:** baseline system follows Tacotron [30] framework consisting of CBHG encoder and auto-regressive decoder [32]. Moreover, we concatenate the lf0, vuv and energy with encoder output.
- **P1:** adopt VAE model with auxiliary speaker classifier $\mathcal{C}$ to extract prosody representation based on BL system.
- **P2:** use SA-WA encoder instead of CBHG encoder based on P1.
- **P3:** the final system we proposed with SA-WA encoder and prosody module both as described in Section 2.

In our implementation, for SA-WA encoder, we use six self-attention blocks, each of which contains two heads of multi-head attention and two linear transformations with 512 and 256 hidden units and the settings of other parts remain the same as [27]. The architecture and hyperparameters of VAE and reference encoder keep the origin configuration [19,22]. The speaker classifier consists of 3 fully connected layers. The decoder adopts an auto-regressive module [32] consisted of prenet, decoder RNN and postnet. In pre-train and fine-tuning stages, the conversion model is trained for 250k steps using batch size of 16, respectively. We use Adam optimizer to optimize our model with learning rate decay, which starts from 0.0002 and decays every 25k steps in decay rate 0.5. The generator G and discriminator D are set to alternate every 5 steps of training, and β is set to 0.1.

3.2. Listening test

We conduct the following listening tests: two mean opinion score (MOS) tests to assess audio quality and speaker similarity respectively, and an AB test to evaluate the performance of source style transfer. We randomly select 15 utterances from each system. Twelve listeners participated in the tests. We highly recommend the readers to listen to our samples[*].

Quality and similarity. For the two MOS tests, listeners are asked to rate the quality and speaker similarity of the converted speech on a 5-point scale. Additionally, the speech from target speaker (TA) is also evaluated with MOS tests. The results of MOS tests are shown in Table 1. It is observed that the speech quality of the five systems is not significantly different. For speaker similarity, BL gets a higher score, and CS achieves a slightly poor result. Since the transfer of source style may affect the timbre and there is no ground truth, these two factors may affect the listeners' judgment. These results show that our proposed method has stability in audio quality and timbre.

Table 1: *MOS results with 95% confidence interval.*

ID	Model	Quality	Speaker Similarity
TA	Target speaker speech	4.41±0.051	4.25±0.042
CS	Comparison system [21]	3.51±0.066	3.51±0.068
BL	Baseline	3.56±0.068	3.65±0.068
P1	+ VAE with classifier $\mathcal{C}$	3.53±0.079	3.61±0.074
P2	+ SA-WA encoder	3.53±0.087	3.58±0.088
P3	+ prosody module	3.54±0.081	3.59±0.083

Source style transfer preference. In the AB test, listeners are asked to choose from paired samples with better style transfer performance. The results are shown in Fig. 3. Comparing with CS, P3 gets a clearly better result. And with the addition of new components and methods, the performance of the source style transfer has also improved. The results validate that both SA-WA encoder and prosody module are beneficial to the perceived performance. Our proposed P3 achieves the best performance.

Figure 3: *Source style transfer preference test results (A:NP:B).*

3.3. Objective Measures

Prosody correlation. To further verify the statistical significance of the expressiveness of each system, we extracted features related to the prosody: frame-level energy and lf0. First, we used 240 samples to calculate the Pearson correlation coefficients between all systems. The higher the Pearson correlation coefficient of the model, the higher the accuracy of the predicted prosodic attributes. As shown in Table 2, P3 gets the highest scores from the perspective of energy and lf0.

Simultaneously, as shown in Fig. 4, we have drawn f0 and energy trajectory diagrams of converted speeches and source audio (SA). Fig. 4 (a) shows that all other systems achieve similar energy curves as SA except for BL. Fig. 4(b) shows a significantly different trend between BL and SA in the curve trend while the other systems keep the same. Among them, P3 achieves a more extensive pitch range and more potent expressiveness. These results above indicate that our proposed system P3 models the source style better.

Table 2: *Correlation in relative energy and lf0.*

	BL	P1	P2	P3
Energy	0.951	0.970	0.970	**0.973**
Lf0	0.665	0.777	0.786	**0.793**

Prosody control. To verify whether lf0 and energy work in the model, we assume that converted speech can be easily controlled through intuitive control of the prosody characteristics. To test this hypothesis, the lf0 and energy of one sentence from test data are multiplied by the coefficients of 0.5, 1 and 1.5, respectively. And we extracted the converted speeches' f0 and energy for visualization. The results are shown in Fig. 5. It can be seen from the figure that as the coefficient increases, the values and fluctuations of the energy and f0 curves become larger. Meanwhile, there is no side effect in the converted speech's timber. This result indicates that lf0 and energy are well-used and can intuitively control the related prosody in our model.

(a) Energy (b) F0
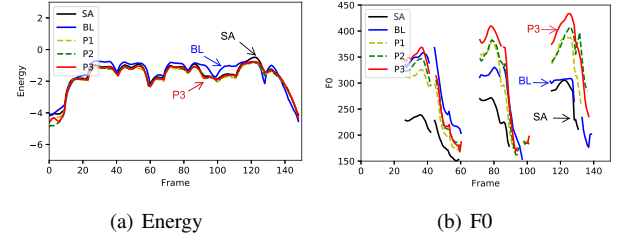
Figure 4: *Energy and f0 values of a converted speech generated by different systems. Please focus on the trend of the curve.*

(a) Energy (b) F0
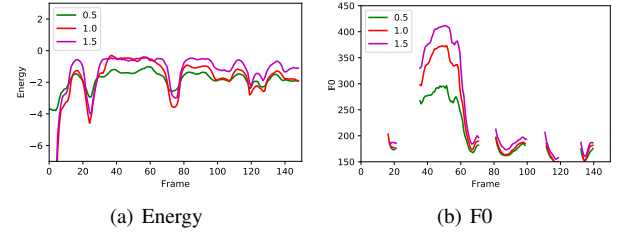
Figure 5: *Intuitive prosody control of energy and f0.*

Visualization of VAE Output. To demonstrate the ability to remove speaker-related information using adversarial training strategy (ADV), we visualize the VAE outputs by t-SNE [33] with and without using ADV in Fig. 6. Thirty utterances from 5 source speakers are selected. The VAE outputs are projected to 2D by t-SNE [33]. Each color represents a speaker. With the help of adversarial training, the outputs from different speakers are uniformly distributed in Fig. 6 (b). This indicates that adversarial training effectively helps VAE to generate speaker-independent outputs.
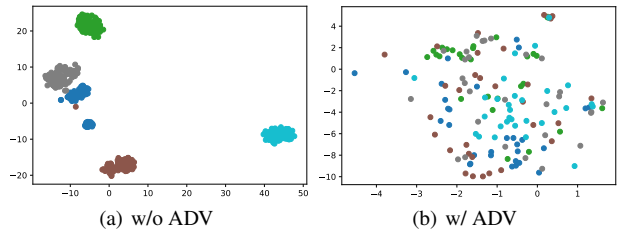

(a) w/o ADV (b) w/ ADV

Figure 6: *VAE output visualization using t-SNE. Each color represents a speaker.*

4. Conclusions

In this study, we propose a hybrid modeling approach for source style transfer in voice conversion based on recognition-synthesis framework, which can transfer the style of source speech to the converted speech. This task is challenging as the speaker's timbre and prosodic style are coupled with each other, and it is difficult to get enough rich and speaker-independent prosody information. Specifically, we first model the prosody in a hybrid manner, which combines implicit and explicit modeling methods. Furthermore, adversarial training is introduced to remove speaker-related information from the extracted prosody. Finally, we use SA-WA encoder to extract comprehensive linguistic content representation. Experimental results show that our proposed system has better abilities in modeling prosody-related acoustic features and controlling synthesized speech prosody.

5. References

[1] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[2] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[3] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[4] E. Godoy, O. Rosec, and T. Chonavel, "Voice Conversion Using Dynamic Frequency Warping With Amplitude Scaling, for Parallel or Nonparallel Corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[5] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Spoken Language Technology Workshop (SLT)*, 2012, pp. 313–317.

[6] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-Based Sparse Representation With Residual Compensation for Voice Conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[7] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1–6.

[8] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[9] X. Tian, E. S. Chng, and H. Li, "A Speaker-Dependent WaveNet for Voice Conversion with Non-Parallel Data," in *Proc. Interspeech*, 2019, pp. 201–205.

[10] Z. Li, B. Tang, X. Yin, Y. Wan, L. Xu, C. Shen, and Z. Ma, "PPG-based singing voice conversion with adversarial representation learning," *ArXiv*, vol. abs/2010.14804, 2020.

[11] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning(ICML)*, 2019, pp. 5210–5219.

[12] J. chieh Chou and H.-Y. Lee, "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization," in *Proc. Interspeech*, 2019, pp. 664–668.

[13] J.-X. Zhang, Z. Ling, and L.-R. Dai, "Non-Parallel Sequence-to-Sequence Voice Conversion With Disentangled Linguistic and Speaker Representations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 540–552, 2020.

[14] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion," in *Porc. Interspeech*, 2016, pp. 2453–2457.

[15] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong, and H. Li, "Exemplar-based sparse representation of timbre and prosody for voice conversion," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5175–5179.

[16] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable Neural Text-to-Speech Synthesis Using Intuitive Prosodic Features," in *Proc. Interspeech*, 2020, pp. 4432–4436.

[17] 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(ICML)*, 2018, pp. 5180–5189.

[18] S. Liu, Y. Cao, S. Kang, N. Hu, X. Liu, D. Su, D. Yu, and H. Meng, "Transferring Source Style in Non-Parallel Voice Conversion," in *Proc. Interspeech*, 2020, pp. 4721–4725.

[19] 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 *International Conference on Machine Learning(ICML)*, 2018, pp. 4693–4702.

[20] J.-X. Zhang, L.-J. Liu, Y.-N. Chen, Y.-J. Hu, Y. Jiang, Z.-H. Ling, and L.-R. Dai, "Voice Conversion by Cascading Automatic Speech Recognition and Text-to-Speech Synthesis with Prosody Transfer," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020, pp. 121–125.

[21] Z. Lian, J. Tao, Z. Wen, B. Liu, Y. Zheng, and R. Zhong, "Towards Fine-Grained Prosody Control for Voice Conversion," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5.

[22] Y.-J. Zhang, S. Pan, L. He, and Z. Ling, "Learning Latent Representations for Style Control and Transfer in End-to-end Speech Synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6945–6949.

[23] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[24] Z. Wang, W. Ge, X. Wang, S. Yang, W. Gan, H. Chen, H. Li, L. Xie, and X. Li, "Accent and Speaker Disentanglement in Many-to-many Voice Conversion," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5.

[25] Q. Wang, W. Rao, S. Sun, L. Xie, C. E. Siong, and H. Li, "Unsupervised Domain Adaptation via Domain Adversarial Training for Speaker Recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4889–4893.

[26] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[27] F. Yang, S. Yang, Q. Wu, Y. Wang, and L. Xie, "Exploiting Deep Sentential Context for Expressive End-to-End Speech Synthesis," in *Proc. Interspeech*, 2020, pp. 3436–3440.

[28] 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(ICML)*, 2018, pp. 2410–2419.

[29] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Trans. Inf. Syst.*, vol. 99-D, pp. 1877–1884, 2016.

[30] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[31] D. Baker, "The db1 Chinese female speech dataset," https://www.data-baker.com/us.html.

[32] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2018.

[33] L. V. D. Maaten and G. E. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.