Interspeech 2022
18-22 September 2022, Incheon, Korea

Cross-speaker Emotion Transfer Based On Prosody Compensation for End-to-End Speech Synthesis

Tao Li[1], *Xinsheng Wang*[2], *Qicong Xie*[1], *Zhichao Wang*[1], *Mingqi Jiang*[3], *Lei Xie*[1*]

[1]Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]School of Software Engineering, Xi'an Jiaotong University, Xi'an, China
[3]Mobvoi AI Lab, Suzhou, China

taoli@npu-aslp.org, w.xinshawn@gmail.com,{xieqicong, zcwang_aslp}@mail.nwpu.edu.cn, mingqi.jiang@mobvoi.com,lxie@nwpu.edu.cn

Abstract

Cross-speaker emotion transfer speech synthesis aims to synthesize emotional speech for a target speaker by transferring the emotion from reference speech recorded by another (source) speaker. In this task, extracting speaker-independent emotion embedding from reference speech plays an important role. However, the emotional information conveyed by such emotion embedding tends to be weakened in the process to squeeze out the source speaker's timbre information. In response to this problem, a prosody compensation module (PCM) is proposed in this paper to compensate for the emotional information loss. Specifically, the PCM tries to obtain speaker-independent emotional information from the intermediate feature of a pre-trained ASR model. To this end, a prosody compensation encoder with global context (GC) blocks is introduced to obtain global emotional information from the ASR model's intermediate feature. Experiments demonstrate that the proposed PCM can effectively compensate the emotion embedding for the emotional information loss, and meanwhile maintain the timbre of the target speaker. Comparisons with state-of-the-art models show that our proposed method presents obvious superiority on the cross-speaker emotion transfer task.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

While the development of attention-based sequence-to-sequence (seq2seq) neural models [1, 2] has brought dramatic improvement for the naturalness of synthetic speech [3, 4, 5], there is still a big gap between synthetic speech and real human speech. Specifically, beyond the naturalness, natural speech produced by human beings is diverse in terms of emotions, which is important paralinguistic information conveyed by human speech. Therefore, how to present appropriate emotions in synthetic speech is vital in immersive human-computer interaction systems, and thus has been drawn much attention recently [6, 7, 8, 9, 10, 11, 12].

A simple case of emotional speech synthesis is the *same-speaker* [13, 14, 12] scenario, in which the speaker identity of the training data is the same as that of synthesized speech. The limitation of this scenario is obvious, i.e., it can only be used for a specific target speaker that has enough desired emotion recordings. Emotional voice conversion [15, 16, 17, 18] is one related task, which aims to convert the emotional state of the utterance from one to another while preserving the linguistic information and speaker identity. In contrast, the *cross-speaker* emotional speech synthesis tries to transfer the emotion from a source speaker who has recorded emotional speech to another (target) speaker, enabling the TTS model to produce emotional speech with the timbre of a speaker who has no emotional speech recordings at all. Reference-based style transfer is the most popular strategy for this cross-speaker case [19, 20, 8, 9, 10]. Reference encoder [20], global style tokens (GST) [8, 10, 21], and variational autoencoder (VAE) [9, 22] are commonly used methods to extract the emotion embedding from the reference with the desired emotion. In these methods, *disentangling* the speaker information from the emotion embedding [10, 23, 24] is important. Otherwise, the speaker information retained in the embedding could contaminate the target timbre.

However, it is challenging to prevent emotional information of the emotion embedding being weakened in the process of removing speaker information, because both the emotion and timbre are related to the prosody, making those reference-based methods have to make a trade-off between the emotion transfer quality and speaker similarity [25]. Specifically, the reference embedding with enough emotion information tends to result in the *speaker leakage*, while further disentangling the speaker information for the embedding could make the emotional expressiveness become weaker in synthetic speech. To face this challenge, in this paper we propose to compensate the emotion embedding for the emotional information loss that caused by the speaker information disentangling, to improve the emotional expressiveness of synthetic speech.

Taking inspiration from [26, 27] that the hidden representation produced by a pre-trained automatic speech recognition (ASR) model retains prosody information but no obvious speaker information, a prosody compensation module (PCM), which takes the ASR model's intermediate feature (AIF) of reference audio as input (as shown in the lower-left corner of Fig. 1), is proposed to compensate the emotion information. The proposed cross-speaker emotional speech synthesis model with prosody compensation, referred to as CSPC, is a Tacotron2 [28] based framework with extended speaker disentangling module (SDM), speaker identity controller, and PCM. To be specific, the SDM is to obtain disentangled speaker-independent emotion embeddings from the reference spectrum, and the PCM is to obtain extra emotion information from AIF to compensate for the information loss in the emotion embedding caused by the disentangling. To effectively extract global

*Corresponding author. This work was supported by the National Key R & D Program of China (2020AAA0108600).

Copyright © 2022 ISCA

10.21437/Interspeech.2022-610

prosody information from AIF, a prosody compensation encoder assisted by global context [29] (GC) blocks is introduced. Extensive experiments have demonstrated the effectiveness of this proposed prosody compensation method on improving the emotional expressive and meanwhile maintaining the voice of the target speaker, and also indicated the effectiveness of the proposed prosody compensation encoder.

In addition to the reference-based method, the most recent work [11] tried to synthesize diverse emotional speech with a label-based approach. To this end, a content-aware prosody prediction module is proposed to predict the prosody from input text with the explicit emotion label. It has shown that this label-based method with extra text-based prosody prediction can achieve comparable performance to the reference-based method. In this paper, in addition to several state-of-the-art reference-based methods, this label-based method will also be compared.

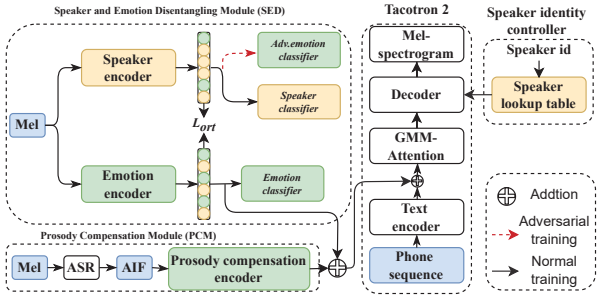

Figure 1: *General architecture of the proposed cross-speaker model.*

2. Proposed Model

The architecture of the proposed model is illustrated in Fig. 1, which shares the skeleton of the modified Tacotron2 [30, 12] while with proposed SDM to obtain speaker-independent emotion embeddings by squeezing out the source timbre information, and PCM to compensate emotion information. The target speaker's identity is provided by the speaker identity controller.

2.1. Speaker Disentangling Module (SDM)

In the reference-based *cross-speaker* emotion transfer speech synthesis method, the emotion embedding obtained from reference audio should be emotionally salient and free of speaker information. To this end, following the basic idea of the speaker disentangling module in [24], the emotion encoder trained with SDM is restrained by an emotion classification constraint to make the obtained emotion embedding discriminative in terms of emotion categories. The corresponding loss is referred to as $\mathcal{L}_{emo}$.

To keep source speaker information out of the emotion embedding, the emotion embedding is made orthogonal to the speaker embedding with an orthogonality loss $\mathcal{L}_{ort}$ [24], which is defined as

$$\mathcal{L}_{ort} = \sum_{i=1}^{n} \left\| s_i^T e_i \right\|_F^2 \quad (1)$$

where $\|\cdot\|_F$ is the Frobenius norm and the e_i is emotion embedding. The speaker embedding s_i is obtained by the speaker encoder in SDM, which is trained with the speaker classification loss $\mathcal{L}_{spk}$ and an emotion classification loss with gradient inversion $\mathcal{L}_{adv_emo}$.

The emotion encoder and speaker encoder share the same input and architecture that consist of six 2D convolutional layers and a GRU layer, and only the last GRU state is taken as the global feature that works as the input of an FC layer to generate the 256-dim embedding.

2.2. Prosody Compensation Module (PCM)

Due to the entanglement of emotional information and speaker information in speech, the emotional information conveyed by the emotion embedding would be weakened after disentangling the speaker-related information, which could lead to insipid expressiveness in synthetic speech. Inspired by the research [26, 27] which shows that the ASR's intermediate feature (AIF) retains the speaker-independent emotional information, PCM is proposed here to compensate the emotion embedding for the emotional information loss using AIF. As illustrated in Fig. 1, the AIF extracted by the encoder of a pre-trained ASR model is used to obtain a global feature via the prosody compensation encoder.

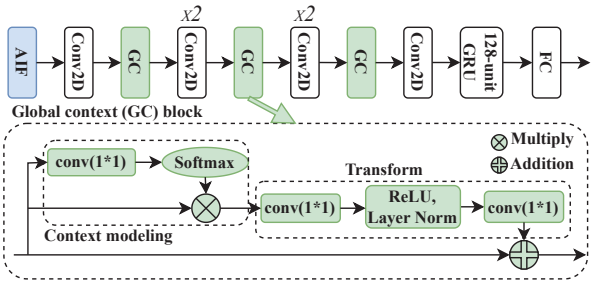

Figure 2: *The architecture of prosody compensation encoder*

While the contextual information contained by the sequence is important for an ASR model, at a specific time step, the prediction of a word or phone is more related to the local context sequence rather than the entire sequence. In contrast, the emotion conveyed by an entire utterance is global information in general. Therefore, how to extract this global information from relatively scattered ASR model's intermediate representations is important. To this end, the global context [29] (GC) block, which shows promising ability on obtaining global information, is used in the prosody compensation encoder. As shown in Fig. 2, this prosody compensation encoder is similar to the emotion encoder in SDM. The difference is that three GC blocks are introduced in this encoder. The GC block [29] consists of global attention pooling, bottleneck transform, and broadcast element-wise addition for context modeling, channel-wise dependency capturing, and feature fusion, respectively, which help the encoder to capture long-range dependency. The prosody compensation embedding produced by this prosody compensation encoder is added to the emotion embedding, which is then added with the tacotron encoder output to work as the conditioning variable for providing emotion information.

2.3. Training and Generation

Here, a Transformer-based ASR model, which is implemented with the WENET[1] [31] toolkit and pre-trained on multi-Chinese corpora containing 2825 hours, is utilized. The encoder of this ASR model produces 256-dim AIF. Parameters of SDM and prosody compensation encoder are optimized together with the skeleton of Tacotron2. The loss function of this skeleton is the

[1]The model is available at `https://github.com/wenet-e2e/wenet`

same as the objective function of Tacotron2 [28], which is denoted as $\mathcal{L}_{taco}$. The final objective function to train the proposed model is given by:

$$\mathcal{L} = \mathcal{L}_{taco} + \mathcal{L}_{emo} + \mathcal{L}_{spk} + \mathcal{L}_{adv_emo} + \alpha\mathcal{L}_{ort}, \quad (2)$$

where $\alpha = 0.1$ is hyper-parameters of $\mathcal{L}_{ort}$.

For the speaker information, we specify the timbre for the model by a trainable speaker look-up table [20] during the training and inference stage, which takes the speaker ID as input and generates a 128-dimensional speaker representation to concatenate with the decoder's input in frame-level. The emotion information is provided by the emotion encoder and prosody compensation module, both of which are with the source speaker's emotional speech as input.

3. EXPERIMENTS

3.1. Experimental Setup

To achieve emotion transfer across speakers, a database that consists of at least two speakers is required, in which one is the source speaker to provide reference emotional speech and another one is the target speaker who only has neutral emotion speech. In this work, the source speaker comes from an individual adult female emotional corpus, referred to as *DB_6*. *DB_6* consists of 5,000 neutral utterances and 2,000 utterances for each emotion category (surprise, happy, sadness, angry, disgust, and fear). The corpus providing the target speaker is *DB_2*, which is another individual adult female corpus that contains about 5,000 samples recorded in a reading style with neutral emotion. Timbres of the source speaker and the target speaker are distinctly different.

During training, all speech samples are down-sampled to 16 kHz, and are then represented by Mel-spectrograms obtained with 50 ms frame length and 12.5 ms frame-shift. WaveRNN [32] is adopted as the neural vocoder to reconstruct waveform from the predicted Mel-spectrograms during inference. Twenty sentences from the target speaker's test set were randomly selected to synthesize speech for each emotion category, resulting in 120 testing samples for each evaluated method.

3.1.1. Evaluation metrics

A good synthetic sample in the emotion transfer task should present the emotion of the reference audio and meanwhile keep the timbre of the target speaker. Here, typical MOS (Mean Opinion Score) tests are conducted to evaluate the emotion similarity and speaker similarity subjectively. In the emotion similarity MOS test, participants are asked to rate synthetic speech based on how well this synthetic speech conveys the same emotion as that conveyed by reference audio, while the speaker similarity MOS test is to ask participants to rate synthetic speech based on how well the synthesized sample sound like uttered by the target speaker. The rating score in both MOS test ranges from *1* to *5* indicating the performance from *bad* to *great*. Besides, another subjective test, named as AB preference test, is adopted in the ablation study to compare samples synthesized by two models. In this AB test, participants are asked to choose one from two samples according to a certain requirement. For each subjective evaluation test, twenty Chinese native speakers participated.

4. Experimental Results

4.1. Comparison with other methods

To evaluate the performance of the proposed method on the *cross-speaker* emotion transfer TTS task, we first compare our method with several state-of-the-art methods [10, 24, 11]. Among those compared methods, multi-reference Tacotron (Multi-R) [10] and cross-speaker emotion transfer (CSET) [24] are reference-based method, and prosody bottleneck (PB) [11] is a recently proposed label-based method.

Table 1: *Comparison with state-of-the-art methods in terms of emotion similarity MOS with confidence intervals of 95%.*

Method	Multi-R [10]	CSET [24]	PB [11]	CSPC
surprise	3.52±0.06	3.73±0.05	3.63±0.04	**3.92±0.06**
happy	3.73±0.05	3.85±0.06	3.76±0.05	**4.01±0.05**
sadness	3.30±0.04	**4.01±0.04**	3.35±0.06	**4.07±0.05**
angry	3.27±0.06	3.69±0.05	3.39±0.04	**3.85±0.05**
disgust	3.21±0.06	3.53±0.04	3.31±0.05	**3.81±0.06**
fear	3.25±0.04	3.78±0.05	3.47±0.05	**3.98±0.05**
average	3.38±0.5	3.76±0.05	3.49±0.04	**3.89±0.05**

Results of emotion similarity MOS test achieved by different models are shown in Table 1, in which higher emotion similarity MOS value means better emotion transfer performance. As shown in this table, the proposed CSPC achieves the best emotion similarity in terms of all emotion categories. Besides, compared with other methods, the proposed CSPC presents a more stable performance in terms of different emotion categories. Specifically, the relative difference between the highest and lowest scores achieved by CSPC in different emotions is 6.8%, while this differences of Multi-R, CSET, and PB are 16.2%, 13.6%, and 13.5% respectively.

Table 2: *Comparison with sate-of-the-art methods in terms of speaker similarity MOS with confidence intervals of 95%.*

Method	Multi-R [10]	CSET [24]	PB [11]	CSPC
surprise	**4.08±0.06**	3.95±0.05	**4.05±0.05**	3.97±0.04
happy	**4.15±0.07**	3.83±0.07	3.90±0.07	3.85±0.04
sadness	**4.02±0.05**	3.82±0.07	3.86±0.08	3.84±0.05
angry	**4.12±0.05**	3.89±0.06	3.93±0.06	3.92±0.06
disgust	4.0±0.06	**4.03±0.04**	3.99±0.07	**4.01±0.06**
fear	**4.15±0.07**	3.98±0.06	4.06±0.04	4.05±0.05
average	**4.08±0.04**	3.91±0.06	3.97±0.05	3.94±0.05

In addition to transferring the emotion from reference to synthesized speech, the timbre of the target speaker should be kept in the synthesized speech. Table 2 shows the performance of different models on keeping the target timbre in synthetic speech, which is evaluated by the speaker similarity MOS test. Higher speaker similarity means better timbre retention. As shown in this table, Multi-R achieves the best speaker similarity MOS in terms of most emotion categories except for *disgust*. However, this good target timbre retention is accompanied by unacceptably terrible emotion transfer as shown in Table 1. Actually, the speaker similarity MOS difference between Multi-R and other methods could be partially caused by the fact that emotional speech could affect participants on the rating of the timbre similarity, because the reference is neutral rather than emotional speech. As for the comparison among CSET, PB, and the proposed CSPC, there is no obvious difference. Considering the obvious superiority of the proposed CSPC on the emotion transfer, the proposed method achieves new state-of-the-art performance on the *cross-speaker* emotion transfer TTS task.

4.2. Ablation study

In the proposed CSPC, emotion information is provided by embeddings with the emotion encoder in SDM and the prosody compensation encoder in PCM. In this section, we would like to show the necessity of utilizing these two embeddings on the emotion transfer. Furthermore, the performance of the GC block in the prosody compensation encoder would also be analyzed in an ablation study.

Table 3: *Performance of different variants of the proposed methods in terms of emotion similarity with confidence intervals of 95%.*

Method	w/o EE	w/o PCE	CSPC
surprise	3.68±0.04	3.77±0.04	**3.92±0.06**
happy	3.75±0.05	3.81±0.05	**4.01±0.05**
sadness	3.44±0.04	3.96±0.04	**4.05±0.05**
angry	3.39±0.06	3.72±0.04	**3.85±0.07**
disgust	3.40±0.07	3.76±0.05	**3.81±0.06**
fear	3.42±0.04	3.71±0.06	**3.98±0.05**
average	3.51±0.04	3.78±0.4	**3.89±0.05**

Table 4: *Performance of different variants of the proposed methods in terms of subjective speaker similarity with confidence intervals of 95%.*

Method	w/o EE	w/o PCE	CSPC
average	**4.01±0.04**	3.96±0.061	3.94±0.054

4.2.1. The necessity of emotion embedding and prosody compensation

The verification of the necessity to use both emotion embedding and prosody compensation embedding is performed in an ablation study, in which the emotion embedding and prosody compensation embedding are dropped respectively. The subjective MOS test results are shown in Table 3 and Table 4, in which "w/o EE" and "w/o PCE" means the emotion embedding and prosody compensation embedding are dropped respectively.

As shown in Table 3, the drop of either emotion embedding (w/o EE) or prosody compensation embedding (w/o PCE) results in an obvious performance drop in terms of all emotion categories. Specifically, the drop of emotion embedding brings the biggest performance drop on the emotion transfer, indicating that the emotion information is mainly conveyed by the emotion embedding. Even if the obvious performance drop appears when the emotion embedding is dropped, the performance of "w/o EE" still outperforms Multi-R and PB (see Table 1), demonstrating that the prosody compensation embedding indeed can convey the emotion information. The performance drop of "w/o PCE" demonstrates that the prosody compensation embedding can provide extra emotion information to the emotion embedding. As for the target timbre retention, by comparing the speaker similarity MOS scores achieved by "w/o PCE" and "CSPC" in Table 4, it can be found that no obvious performance drop appears when the prosody compensation embedding is added. All those results indicate that the ASR model's intermediate feature (AIF) can effectively provide prosodic information to enable the proposed prosody compensation method to further improve the emotion transfer performance and meanwhile maintain the target speaker's timbre.

4.2.2. The effectiveness of GC block

The effectiveness of the emotion compensation embedding has been verified in Section 4.2.1. Here, we would like to show the importance of the GC block in the proposed emotion compensation encoder. To this end, an AB preference test between the proposed method and a variant of the proposed method, named as "w/o GC", in which the GC block is dropped from the emotion compensation encoder (See Fig. 2) is conducted.

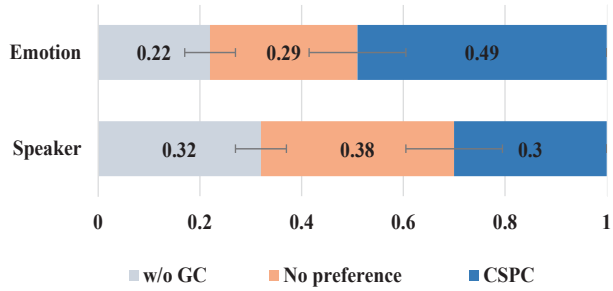

Figure 3: *Emotion similarity and speaker similarity AB preference test for the ablation study of GC block with confidence intervals of 95%.*

The results are shown in Fig. 3, in which the global voting percents over different emotion categories are presented on the demo page[2] (Additional demos are also provided which transfer emotion to the new target speaker who only has 200 sentences through simple fine-tuning.). As can be seen from this figure, in terms of the emotion similarity, CSPC significantly outperforms the variant without the GC block (p-value = 0.042). As for the speaker similarity, dropping the GC block brings a slight increase which is not significant (p-value = 0.063). These results demonstrate that the proposed prosody compensation encoder is well designed, and benefits from the GC block on capturing the long-range dependency emotion information, which can better obtain the global emotion-related information from AIF.

5. Conclusion

In this paper, a prosody compensation method is proposed to improve the emotion transfer in the cross-speaker emotion transfer task. Inspired by the fact that the intermediate feature obtained by a pre-trained ASR model could contain prosody information, the prosody embedding extracted by the proposed prosody compensation module (PCM), which takes the ASR model's intermediate feature (AIF) as input, is used to provide extra information to the disentangled emotion embedding produced by a speaker disentangling module (SDM). Extensive experiments have demonstrated that the proposed prosody compensation method can successfully provide extra emotion information to the emotion embedding, for which the lost emotion information due to the speaker-emotion disentangling process could be well compensated. Furthermore, the ablation study shows that the GC block plays an important role in obtaining the prosody embedding, indicating the good design of the proposed prosody embedding encoder in PCM.

[2]Audio samples and detailed results in terms of each emotion category can be found on the project page `https://silyfox.github.io/cspc/`

6. References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2015.

[2] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, and S. Bengio, "Tacotron: Towards end-to-end speech synthesis," *INTERSPEECH*, 2017.

[4] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *NeurIPS*, 2019.

[5] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "Durian: Duration informed attention network for multimodal synthesis," *arXiv preprint arXiv:1909.01700*, 2019.

[6] H. jin Choi, S. Park, J. Park, and M. Hahn, "Multi-speaker emotional acoustic modeling for CNN-based speech synthesis," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6950–6954, 2019.

[7] S.-Y. Um, S. Oh, K. Byun, I. Jang, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[8] 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," *in Proc. ICML, 2018, pp. 5180–5189.*

[9] 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," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6945–6949, 2019.

[10] Y. Bian, C. Chen, Y. Kang, and Z. Pan, "Multi-reference Tacotron by intercross training for style disentangling,transfer and control in speech synthesis," *arXiv preprint arXiv:1904.02373,*, 2019.

[11] S. Pan and L. He, "Cross-speaker style transfer with prosody bottleneck in neural speech synthesis," in *Interspeech*, 2021.

[12] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5, 2021.

[13] Y. Lee, A. Rabiee, and S. Y. Lee, "Emotional end-to-end neural speech synthesizer," *arXiv preprint arXiv:1711.05447,*, 2017.

[14] R. Li, Z. Wu, Y. Huang, J. Jia, H. Meng, and L. Cai, "Emphatic speech generation with conditioned input layer and bidirectional LSTMS for expressive speech synthesis," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5129–5133, 2018.

[15] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," *ArXiv*, vol. abs/2002.00198, 2020.

[16] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *ArXiv*, vol. abs/2105.14762, 2021.

[17] Y. Cao, Z. Liu, M. Chen, J. Ma, S. Wang, and J. Xiao, "Non-parallel emotional speech conversion using vae-gan," in *INTERSPEECH*, 2020.

[18] H. Choi and M. Hahn, "Sequence-to-sequence emotional voice conversion with strength control," *IEEE Access*, vol. 9, pp. 42 674–42 687, 2021.

[19] A. Gibiansky, S. Ö. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *NIPS*, 2017.

[20] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. A. J. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," *ArXiv*, vol. abs/1803.09047, 2018.

[21] O. Kwon, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis based on style embedded Tacotron2 framework," *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pp. 1–4, 2019.

[22] A. Kulkarni, V. Colotte, and D. Jouvet, "Improving latent representation for end to end multispeaker expressive text to speech system," ArXiv,2020.

[23] M. Whitehill, S. Ma, D. McDuff, and Y. Song, "Multi-reference neural TTS stylization with adversarial cycle consistency," *ArXiv*, vol. abs/1910.11958, 2020.

[24] T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie, "Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1448–1460, 2022.

[25] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Saez-Trigueros, and T. Drugman, "Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech," in *INTERSPEECH*, 2020.

[26] 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," *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5, 2021.

[27] Z. Wang, X. Zhou, F. Yang, T. Li, H. Du, L. Xie, W. Gan, H. Chen, and H. Li, "Enriching source style transfer in recognition-synthesis based non-parallel voice conversion," in *Interspeech*, 2021.

[28] 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 *Proc. ICASSP*. IEEE, 2018, pp. 4779–4783.

[29] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 1971–1980, 2019.

[30] F. Yang, S. Yang, P. Zhu, P. Yan, and L. Xie, "Improving Mandarin end-to-end speech synthesis by self-attention and learnable Gaussian bias," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 208–213, 2019.

[31] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech*, Brno, Czech Republic, 2021.

[32] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *ArXiv*, vol. abs/1802.08435, 2018.