PAVITS: EXPLORING PROSODY-AWARE VITS FOR END-TO-END EMOTIONAL VOICE CONVERSION

Tianhua Qi[1,2], *Wenming Zheng*[1,2*], *Cheng Lu*[1,2], *Yuan Zong*[1,2], *Hailun Lian*[1]

[1]Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education, Nanjing 210096, China
[2]School of Biological Science and Medical Engineering, Southeast University, China
{qitianhua, wenming_zheng*, cheng.lu, xhzongyuan, lianhailun}@seu.edu.cn

ABSTRACT

In this paper, we propose Prosody-aware VITS (PAVITS) for emotional voice conversion (EVC), aiming to achieve two major objectives of EVC: high content naturalness and high emotional naturalness, which are crucial for meeting the demands of human perception. To improve the content naturalness of converted audio, we have developed an end-to-end EVC architecture inspired by the high audio quality of VITS. By seamlessly integrating an acoustic converter and vocoder, we effectively address the common issue of mismatch between emotional prosody training and run-time conversion that is prevalent in existing EVC models. To further enhance the emotional naturalness, we introduce an emotion descriptor to model the subtle prosody variations of different speech emotions. Additionally, we propose a prosody predictor, which predicts prosody features from text based on the provided emotion label. Notably, we introduce a prosody alignment loss to establish a connection between latent prosody features from two distinct modalities, ensuring effective training. Experimental results show that the performance of PAVITS is superior to the state-of-the-art EVC methods. Speech Samples are available at `https://jeremychee4.github.io/pavits4EVC/`.

Index Terms— Emotional voice conversion, end-to-end model, prosody, emotional speech, multi-task learning

1. INTRODUCTION

Emotional voice conversion (EVC) endeavors to transform the state of a spoken utterance from one emotion to another, while preserving the linguistic content and speaker identity [1]. It brings the capability to facilitate emotional communication between individuals [2], enhancing the user experience in human-computer interaction [3], and even achieving a seamless integration of human presence within the virtual world [4].

There are two distinct challenges in EVC: one is low content naturalness, and the other is that the converted audio lacks the richness of emotion compared to human voice [1]. Previous studies were focused on frame-based solutions, such as CycleGAN [5] and StarGAN [6, 7]. However, due to the fixed-length nature and poor training stability, the naturalness of converted audio is quite low to apply in practice. To address this challenge, autoencoder-based [8, 9] especially for sequence-to-sequence (seq2seq) [10, 11] frameworks raise much interests for its variable-length speech generation. It achieves an acceptable naturalness through the joint training with Text-to-speech (TTS) [12], which is used to capture linguistic information and avoid mispronunciation as well as skipping-words. Since speech emotion is inherently supra-segmental [13], it is difficult to learn emotional representation from the spectrogram. To tackle this, various pretraining methods, such as leveraging speech emotion recognition (SER) model [14] and 2-stage training strategy [15], are introduced to extract emotional feature for EVC system.

Despite these works have achieved great success in EVC, the converted audio still falls short in meeting human's perceptual needs, which implies that these two challenges still remain to be effectively addressed. Remarkably, current EVC models generally operate in a cascade manner, i.e., the acoustic converter and the vocoder [1, 5, 7, 8], resulting in a mismatch between emotional prosody training and run-time conversion, ultimately leading to a degradation in audio quality, which is vital to evaluate content naturalness and impacts the perceptual experience of emotional utterance. However, there is no EVC model that attempt to bridge this gap, let alone models that aim to capture prosody variations at a finer granularity. To handle the similar issue, multiple solutions have been explored in TTS, including FastSpeech2s [16], EATS [17], VITS [18, 19], etc., seeking to alleviate the mismatch between acoustic feature generation and waveform reconstruction by integrating these two stages together.

In this paper, inspired by the high audio quality of VITS [18], we propose Prosody-aware VITS (PAVITS) for EVC, a novel end-to-end system with implicit prosody modeling to enhance content naturalness and emotional naturalness. To our best knowledge, PAVITS is the first EVC method in solving the mismatch between acoustic feature conversion and waveform reconstruction. Compared to original VITS, our approach involves several key innovations. In order to improve content naturalness with speech quality, we build upon VITS to solve the two-stage mismatch in EVC, and apply multi-task learning since TTS can significantly reduce the mispronunciation. To enhance emotional naturalness, we introduce an emotion descriptor to capture prosody differences associated with different emotional states in speech. By utilizing Valence-Arousal-Dominance values as condition, emotional representation at utterance-level is learned. Latent code is further refined by a prosody integrator, which incorporates with speaker identity and linguistic content to model finer-grained prosody variations. Then frame-level prosody features are obtained from normalizing flow. We also introduce a prosody predictor that leverages emotion labels and phoneme-level text embedding to predict frame-level emotional prosody features. Finally, we devise a prosody alignment loss to connect two modalities, aligning prosody features obtained from audio and text, respectively.

This work was supported in part by the National Key R & D Project under the Grant 2022YFC2405600, in part by the NSFC under the Grant U2003207 and 61921004, and in part by the Jiangsu Frontier Technology Basic Research Project under the Grant BK20192004.

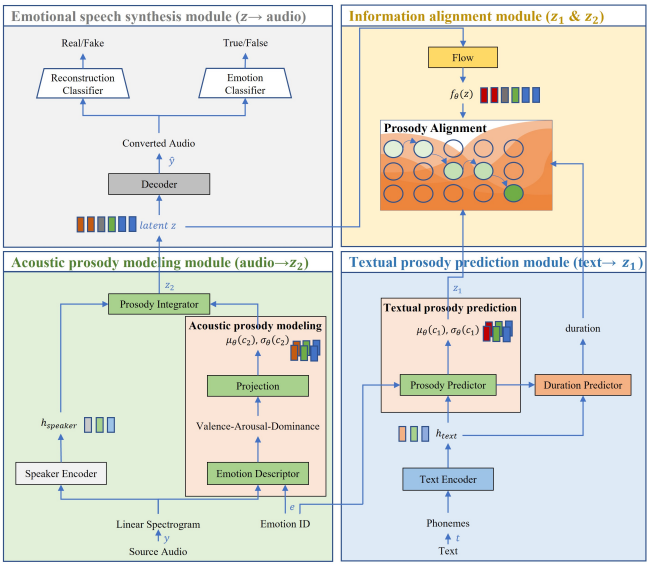

Fig. 1. Architecture of PAVITS.

2. PROPOSED METHOD

As shown in Figure 1, inspired by VITS [18], the proposed model is constructed based on conditional variational autoencoder (CVAE), consisting of four parts: a textual prosody prediction module, an acoustic prosody modeling module, an information alignment module, and an emotional speech synthesis module.

The textual prosody prediction (TPP) module predicts the prior distribution $p(z_1 \mid c_1)$ as:

$$z_1 = TPP(c_1) \sim p(z_1 \mid c_1) \tag{1}$$

where c_1 including text t and emotion label e.

The acoustic prosody modeling (APM) module disentangles emotional features with intricate prosody variation, speaker identity, and linguistic content from the source audio given emotion label, forming the posterior distribution $q(z_2 \mid c_2)$ as:

$$z_2 = APM(c_2) \sim q(z_2 \mid c_2) \tag{2}$$

where c_2 including audio y and emotion label e.

The information alignment module facilitates the alignment of text and speech, as well as the alignment of textual and acoustic prosody representations. In emotional speech synthesis (ESS) module, the decoder reconstructs waveform $\hat{y}$ according to latent representation z.

$$\hat{y} = Decoder(z) \sim p(y \mid z) \tag{3}$$

where z comes from z_1 or z_2.

While the proposed model can perform both EVC and emotional TTS after training, EVC will be the main focus of this paper. In the following, we will introduce the details of the four modules.

2.1. Textual prosody prediction module

Given condition c_1 including text t and emotion label e, the textual prosody prediction module provides the prior distribution $p(z_1 \mid c_1)$ of CVAE. The text encoder takes phonemes as input and extracts linguistic information h_{text} at first. Considering the extensive prosody variation associated with each phoneme, we employ a prosody predictor to extend the representation to frame-level and predict the prosody variation (a fine-grained prior normal distribution with mean μ_θ and variance σ_θ generated by a normalizing flow f_θ) based on emotion label.

$$p(z_1 \mid c_1) = N(f_\theta(z_1); \mu_\theta(c_1); \sigma_\theta(c_1)) \left| \det \frac{\partial f_\theta(z_1)}{\partial z} \right| \tag{4}$$

Text Encoder: Since the training process is constrained by the volume of textual content within parallel datasets, we initially convert text or characters into a phoneme sequence as a preprocessing step to maximize the utility of the available data, resulting in improved compatibility with the acoustic prosody modeling module. Similar to VITS [18], text encoder comprises multiple Feed-Forward Transformer (FFT) blocks with a linear projection layer for representing linguistic information.

Prosody Predictor: Prosody predictor leverages phoneme-level linguistic information extracted by the text encoder to anticipate frame-level prosody variation given discrete emotion label. It has been observed that simply increasing the depth of stacked flow does not yield satisfactory emotional prosody variations, unlike the prosody predictor. Therefore, the inclusion of the prosody predictor guarantees a continuous enhancement in prosody modeling for both the TPP and APM modules. The prosody predictor comprises multiple one-dimensional convolution layers and a linear projection layer. Furthermore, we integrate predicted emotional prosody information with linguistic information as input for the duration predictor, which significantly benefits the modeling of emotional speech duration.

2.2. Acoustic prosody modeling module

The acoustic prosody modeling module provides emotional features with fine-grained prosody variation based on dimensional emotion representation, i.e., Valence-Arousal-Dominance values. Speaker identity and speech content information are also disentangled from the source audio and then complete feature fusion through the prosody integrator as the posterior distribution $q(z_2 \mid c_2)$.

$$q(z_2 \mid c_2) = N(f_\theta(z_2); \mu_\theta(c_2); \sigma_\theta(c_2)) \tag{5}$$

Speaker encoder: Considering the APM module's increased focus on understanding emotional prosody more thoroughly compared to previous models, it's apparent that speaker characteristics could unintentionally be overlooked during conversion. Recognizing the critical role of fundamental frequency (F0) in speaker modeling [20], we augment the F0 predictor of [21] by adding multiple one-dimensional convolutional layers and a linear layer to construct the speaker encoder, which tackles the issue effectively.

Emotion descriptor: To enhance PAVITS's emotional naturalness, we employ a specific SER system rooted in Russell's circumplex theory [22] to predict dimensional emotion representation, encompassing Valence-Arousal-Dominance values as a conditional input. This input guides the capture of nuanced prosody variations, which ensures that while satisfying human perception of emotions at utterance-level, natural prosody variations are retained from segment-level down to frame-level, preserving intricate details. It consists of a SER module [23] and a linear projection layer.

Prosody Integrator: The prosody integrator incorporates a combination of speaker identity attributes, emotional prosody characteristics, and intrinsic content properties extracted from the linear spectrogram. It is constructed using multiple convolution layers, Wavenet residual blocks, and a linear projection layer.

2.3. Information alignment module

In VITS [18], the existing alignment mechanism, which is called Monotonic Alignment Search (MAS), solely relies on textual and acoustic features from parallel datasets. Thus, it is insufficient in capturing emotional prosody nuances, hindering effective linkage between the TPP and APM modules. To overcome this limitation, we propose an additional prosody alignment loss function based on Kullback-Leibler divergence, to facilitate joint training for frame-level prosody modeling across the TPP and APM modules, with the goal of enhancing prosody information integration and synchronization within our model.

$$L_{psd} = D_{KL}\left(q\left(z_2 \mid c_2\right) \| \, p\left(z_1 \mid c_1\right)\right) \tag{6}$$

2.4. Emotional speech synthesis module

In the emotional speech synthesis module, the decoder generates a waveform based on latent z, employing adversarial learning to continuously enhance naturalness in both content and emotion. To improve the naturalness of content, $L_{\text{recon\_cls}}$ minimizes the L1 distance between predicted and target spectrograms, $L_{\text{recon\_fm}}$ minimizes the L1 distance between feature maps extracted from intermediate layers in each discriminator, aimed at enhancing training stability. Since the former predominantly influences the early-to-mid stage, while the latter assumes a more prominent role in mid-to-late stage, we introduce two coefficients to balance their contributions as follows.

$$L_{recon} = \gamma L_{recon\_cls} + \beta L_{recon\_fm}(G) \tag{7}$$

To enhance the perception of emotions, $L_{\text{emo\_cls}}$ represents the loss function for emotional classification, while $L_{\text{emo\_fm}}$ denotes the loss associated with feature mapping for emotion discrimination.

$$L_{emo} = L_{emo\_cls} + L_{emo\_fm}(G) \tag{8}$$

2.5. Final loss

By combining CVAE with adversarial training, we formulate the overall loss function as follows:

$$L = L_{recon} + L_{adv}(G) + L_{emo} + L_{psd} + L_{F0} + L_{dur} \tag{9}$$

$$L(D) = L_{adv}(D) \tag{10}$$

where $L_{adv}(G)$ and $L_{adv}(D)$ represent the adversarial loss for the Generator and Discriminator respectively, L_{F0} minimizes the L2 distance between the predicted F0 and corresponding ground truth, L_{dur} minimizes the L2 distance between the predicted duration and ground truth which is obtained through estimated alignment.

2.6. Run-time conversion

At runtime, there are two converting methods: a fixed-length approach (Audio-z_2-Audio, named PAVITS-FL) and a variable-length approach (Audio-Text-z_1-Audio, named PAVITS-VL). The former uses APM module for latent z prediction from audio, ensuring robustness as it remains unaffected by text encoding, but is constrained by a fixed spectrum length due to Dynamic Time Warping (DTW) limitations. The latter employs TPP module to predict latent z from corresponding text obtained through automatic speech recognition (ASR) techinique, which is not bound by duration modeling and offers greater naturalness. Finally, the ESS module's decoder takes latent z (either z_1 or z_2) as input and synthesizes the converted waveform without a separate vocoder.

Table 1. A comparison of MCD [dB] values.

Model	MCD [dB]			
	Neu-Ang	Neu-Hap	Neu-Sad	Neu-Sur
CycleGAN	4.41	4.24	4.32	5.68
StarGAN	4.52	4.46	4.31	5.79
Seq2seq-WA2	3.73	3.72	3.77	5.60
VITS	3.68	3.70	3.69	5.41
PAVITS-FL (proposed)	**3.42**	3.63	3.40	4.61
PAVITS-VL (proposed)	3.58	**3.62**	**2.98**	**3.96**

3. EXPERIMENTS

3.1. Dataset

We perform emotional conversion on a Mandarin corpus belonged to Emotional Speech Dataset (ESD) [24] from neutral to angry, happy, sad, and surprise, denoted as *Neu-Ang*, *Neu-Hap*, *Neu-Sad*, *Neu-Sur* respectively. For each emotion pair, we use 300 utterances for training, 30 utterances for evaluation, and 20 utterances for test. The total duration of training data is around *80 minutes* (*16 minutes* per emotion category), which is absolutely small compared to others.

3.2. Experimental Setup

We train the following models for comparison.

- CycleGAN [25] (*baseline*): CycleGAN-based EVC model with WORLD vocoder.
- StarGAN [26] (*baseline*): StarGAN-based EVC model with WORLD vocoder.
- Seq2seq-WA2 [15] (*baseline*): Seq2seq-based EVC model employing 2-stage training strategy with WaveRNN vocoder.
- VITS [18] (*baseline*): EVC model constructed by original VITS, operating independently in both fixed-length and variable-length, take the average as the result.
- PAVITS-FL (*proposed*): the proposed model based on VITS, incorporates all the contributions outlined in the paper, but operate within a fixed-length framework.
- PAVITS-VL (*proposed*): the proposed model based on VITS, incorporates all the contributions outlined in the paper, but operate within a variable-length framework leveraging ASR to obtain text from source audio.

3.3. Results & Discussion

Mel-cepstral distortion (MCD) was calculated for objective evaluation, as depicted in Table 1. In terms of subjective evaluation, Mean Opinion Score (MOS) tests were conducted to appraise both the quality and naturalness of speech as shown in Table 2. The naturalness score was derived by averaging the scores for content naturalness and emotional prosody naturalness, as rated by 24 participants, each of whom assessed a total of 148 utterances. We further report emotional similarity results between converted audio and human voice to gauge emotional naturalness as illustrated in Figure 2.

Through the above-mentioned metrics, it is obvious that the proposed PAVITS achieves competitive performance on both objective and subjective evaluation. From the perspective of objective MCD and subjective MOS, both original VITS and our proposed PAVITS models always outperform other models with traditional vocoder or neural vocoder, which proves that the integration of neural acoustic

Table 2. Experimental results in terms of subjective mean opinion score (MOS)

EVC Model	MOS							
	Speech Quality				Naturalness			
	Neu-Ang	Neu-Hap	Neu-Sad	Neu-Sur	Neu-Ang	Neu-Hap	Neu-Sad	Neu-Sur
CycleGAN	3.91±0.19	4.04±0.16	3.95±0.13	3.84±0.12	3.83±0.19	4.01±0.21	3.86±0.20	3.90±0.14
StarGAN	3.53±0.10	3.50±0.12	3.46±0.14	3.49±0.07	3.56±0.20	3.61±0.14	3.71±0.18	3.70±0.17
Seq2seq-WA2	3.95±0.14	4.03±0.24	4.14±0.29	4.03±0.16	3.72±0.14	3.67±0.15	3.72±0.17	3.89±0.20
VITS	4.49±0.06	4.40±0.13	4.55±0.12	4.51±0.06	4.00±0.19	4.15±0.12	4.23±0.20	4.26±0.15
PAVITS-FL (proposed)	4.62±0.04	4.62±0.04	**4.64±0.04**	**4.66±0.02**	4.25±0.19	4.44±0.09	4.48±0.07	4.40±0.13
PAVITS-VL (proposed)	**4.72±0.02**	**4.72±0.01**	4.63±0.03	4.66±0.03	**4.39±0.14**	**4.60±0.11**	**4.59±0.05**	**4.61±0.10**
Ground Truth	4.78±0.02	4.81±0.01	4.82±0.01	4.86±0.01	4.71±0.06	4.78±0.05	4.83±0.02	4.80±0.04

Table 3. Ablation study with MOS test

EVC Model	Speech Quality	Naturalness
PAVITS (proposed)	**4.67±0.04**	**4.60±0.07**
w/o Prosody Predictor	4.48±0.10	4.16±0.13
w/o Prosody Alignment	4.38±0.05	4.08±0.10
w/o Prosody Integrator	4.56±0.09	4.37±0.17

converter and vocoder is suitable for EVC task to enhance speech quality and naturalness. It is worth noting that even in the case of the fixed-length PAVITS-FL model, there is a reduction of over 0.4 in MCD when compared to the variable-length seq2seq model and the original VITS model. Furthermore, there has been an enhancement of 0.6 and 0.2 in MOS, respectively. To some extent, it reflects how human tend to be influenced by audio quality when assessing model naturalness, especially when there are significant differences in quality being compared.

As depicted in Figure 2, our proposed PAVITS-VL (variable-length) model aligns more closely with human perception in the converted audio, which attributed to the model's capacity for fine-grained granularity in modeling speech emotion, incorporating implicit prosody cues. To further show the effectiveness of our method, we visualize the spectrogram of testing clips, as exemplified in Figure 3. It is readily apparent that the spectrogram converted by PAVITS exhibits finer details in prosody variations within the pertinent frequency bands, while simultaneously preserving descriptive information for other frequency bands. Consequently, the audio generated by PAVITS possesses a prosody naturalness and emotional accuracy that closely approximates the ground truth spectrogram.

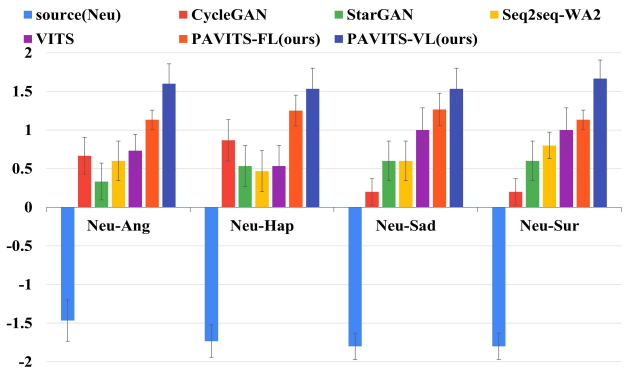

Fig. 2. Emotional similarity test with 95% confidence interval following [15].

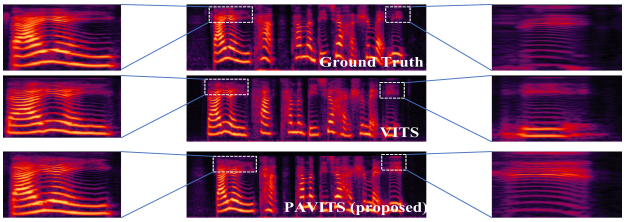

Fig. 3. Spectrogram of a testing clip (happy), from top to bottom are ground truth, converted by original VITS, and proposed PAVITS.

3.4. Ablation Study

We further conduct an ablation study to validate different contributions. We remove prosody predictor, prosody alignment, and prosody integrator in turn and let the subjects evaluate quality and naturalness of converted audio. From Table 3, we can see that all scores are degraded with the removal of different components. When remove prosody predictor, the speech quality does not undergo significant changes, as the original VITS primarily relies on textual features as input. However, a significant decrease in naturalness is observed, attributed to the loss of explicit emotion label for TPP module as a conditioning factor. This highlights the importance of aligning with APM module on the basis of information asymmetry, which reflects the ingenious design of prosody modeling structure. Note that the performance of PAVITS is worse than VITS after deleting prosody alignment, it might be attributed the fact that latent prosody representations are not constrained during training, which damages the original MAS mechanism present in VITS. To further show the contribution from the prosody integrator, we replace it with a simple concatenation. Both speech quality and naturalness show a slight decrease, indicating that utilizing prosody integrator for information fusion is quite effective for APM module.

4. CONCLUSION

In this paper, we propose Prosody-aware VITS (PAVITS) for emotional voice conversion (EVC). By integrating acoustic prosody modeling (APM) module with textual prosody prediction (TPP) module through prosody alignment, the fine-grained emotional prosody features across various scales of emotional speech can be learned effectively. Experimental results on ESD corpus demonstrate the superiority of our proposed PAVITS for content naturalness and emotional naturalness, even when dealing with limited data scenarios. In the future, we will explore the controllable emotional prosody modeling to allow better interpretability of EVC.

5. REFERENCES

[1] Kun Zhou, Berrak Sisman, Rajib Rana, Björn W. Schuller, and Haizhou Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 31–48, jan 2023.

[2] Chunhui Lu, Xue Wen, Ruolan Liu, and Xiao Chen, "Multi-speaker emotional speech synthesis with fine-grained prosody modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 5729–5733.

[3] Rajdeep Chatterjee, Saptarshi Mazumdar, R Simon Sherratt, Rohit Halder, Tanmoy Maitra, and Debasis Giri, "Real-time speech emotion analysis for smart home assistants," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 1, pp. 68–76, 2021.

[4] John David N Dionisio, William G Burns Iii, and Richard Gilbert, "3d virtual worlds and the metaverse: Current status and future possibilities," *ACM Computing Surveys (CSUR)*, vol. 45, no. 3, pp. 1–38, 2013.

[5] Changzeng Fu, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro, "An improved cyclegan-based emotional voice conversion model by augmenting temporal dependency with a transformer," *Speech Communication*, vol. 144, pp. 110–121, 2022.

[6] Zongyang Du, Berrak Sisman, Kun Zhou, and Haizhou Li, "Expressive voice conversion: A joint framework for speaker identity and emotional style transfer," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 594–601.

[7] Xiangheng He, Junjie Chen, Georgios Rizos, and Björn W Schuller, "An improved stargan for emotional voice conversion: Enhancing voice quality and data augmentation," *arXiv preprint arXiv:2107.08361*, 2021.

[8] Xunquan Chen, Xuexin Xu, Jinhui Chen, Zhizhong Zhang, Tetsuya Takiguchi, and Edwin R Hancock, "Speaker-independent emotional voice conversion via disentangled representations," *IEEE Transactions on Multimedia*, 2022.

[9] Wenhuan Lu, Xinyue Zhao, Na Guo, Yongwei Li, Jianguo Wei, Jianhua Tao, and Jianwu Dang, "One-shot emotional voice conversion based on feature separation," *Speech Communication*, vol. 143, pp. 1–9, 2022.

[10] Zijiang Yang, Xin Jing, Andreas Triantafyllopoulos, Meishu Song, Ilhan Aslan, and Björn W Schuller, "An overview & analysis of sequence-to-sequence emotional voice conversion," 2022.

[11] Zhiyuan Zhao, Jingjun Liang, Zehong Zheng, Linhuang Yan, Zhiyong Yang, Wan Ding, and Dongyan Huang, "Improving model stability and training efficiency in fast, high quality expressive voice conversion system," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 75–79.

[12] Tae-Ho Kim, Sungjae Cho, Shinkook Choi, Sejik Park, and Soo-Young Lee, "Emotional voice conversion using multitask learning with text-to-speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 7774–7778.

[13] Srividya Tirunellai Rajamani, Kumar T Rajamani, Adria Mallol-Ragolta, Shuo Liu, and Björn Schuller, "A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 6294–6298.

[14] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 920–924.

[15] Kun Zhou, Berrak Sisman, and Haizhou Li, "Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training," *arXiv preprint arXiv:2103.16809*, 2021.

[16] 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 *International Conference on Learning Representations*, 2020.

[17] Jeff Donahue, Sander Dieleman, Mikolaj Binkowski, Erich Elsen, and Karen Simonyan, "End-to-end adversarial text-to-speech," in *International Conference on Learning Representations*, 2020.

[18] Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[19] Yuma Shirahata, Ryuichi Yamamoto, Eunwoo Song, Ryo Terashima, Jae-Min Kim, and Kentaro Tachibana, "Period vits: Variational inference with explicit pitch modeling for end-to-end emotional speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

[20] Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 582–596, 2009.

[21] Yongmao Zhang, Jian Cong, Heyang Xue, Lei Xie, Pengcheng Zhu, and Mengxiao Bi, "Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 7237–7241.

[22] James A Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.

[23] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[24] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[25] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," 2019.

[26] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha, "Stargan v2: Diverse image synthesis for multiple domains," 2020.