Interspeech 2024
1-5 September 2024, Kos, Greece

Towards Realistic Emotional Voice Conversion using Controllable Emotional Intensity

Tianhua Qi[1,2], *Shiyan Wang*[1,2], *Cheng Lu*[1,2], *Yan Zhao*[1], *Yuan Zong*[1,2], *Wenming Zheng*[*1,2]

[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@seu.edu.cn, xhzongyuan@seu.edu.cn, wenming_zheng*@seu.edu.cn

Abstract

Realistic emotional voice conversion (EVC) aims to enhance emotional diversity of converted audios, making the synthesized voices more authentic and natural. To this end, we propose Emotional Intensity-aware Network (EINet), dynamically adjusting intonation and rhythm by incorporating controllable emotional intensity. To better capture nuances in emotional intensity, we go beyond mere distance measurements among acoustic features. Instead, an emotion evaluator is utilized to precisely quantify speaker's emotional state. By employing an intensity mapper, intensity pseudo-labels are obtained to bridge the gap between emotional speech intensity modeling and run-time conversion. To ensure high speech quality while retaining controllability, an emotion renderer is used for combining linguistic features smoothly with manipulated emotional features at frame level. Furthermore, we employ a duration predictor to facilitate adaptive prediction of rhythm changes condition on specifying intensity value. Experimental results show EINet's superior performance in naturalness and diversity of emotional expression compared to state-of-the-art EVC methods.

Index Terms: emotional voice conversion, emotional intensity modeling, fine-grained control, realistic speech synthesis

1. Introduction

"Everything I read I try to figure out: what it really means, what it's really saying."

–Richard P. Feynman

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 holds the promise of fostering more profound emotional communication between individuals [2], elevating user experience in human-machine interactions [3], as well as creating a more immersive and resonant virtual experience [4].

Current EVC systems are predominantly constructed based on autoencoder [5, 6, 7, 8] especially for sequence-to-sequence (seq2seq) [9, 10] frameworks, with significant strides in speech quality. However, the converted audio lacks emotional diversity, which is a critical aspect for achieving realistic speech synthesis. Therefore, incorporating an intensity control module into typical EVC framework has become a primary research focus to facilitate manipulation of emotional expression, consequently addressing one-to-many problem in a controllable manner.

For instance, Emovox [11] is constructed based on Seq2seq-EVC [12], leveraging the relative attribute ranking (RAR) [13] metric to measure relative difference among acoustic features such as pitch frequency and frame energy, between emotional and non-emotional speech samples. Additionally, intensity pseudo-labels are generated to address the absence of explicit annotations in emotional corpora [14]. Similarly, AINN [15] is built upon EmotionalStarGAN [16], incorporating contrastive learning to construct positive and negative pairs. The calculation of intensity pseudo-labels is also employed to control emotion transformation by explicitly specifying an intuitive intensity value.

Despite the great success achieved by intensity control approaches in EVC, the converted vocal expressiveness still falls short of meeting human perceptual expectations, particularly in terms of naturalness and diversity. This deficiency can be attributed to the prevalent utilization of intensity modeling methods that solely rely on measuring the differences in acoustic features [11, 15, 17, 18]. This dependency overlooks inherent emotional fluctuations of speaker, leading to a mismatch between emotional intensity modeling and run-time conversion, which poses a substantial obstacle to synthesize authentic voices.

The dimensional representation method allows for a more accurate portrayal of the distinctions between emotional states, drawing inspiration from the circumplex theory [19]. As proposed by [20], within the 2-dimensional VA-space formed by valence and arousal, the wedge area formed by these dimensions can be utilized to gauge emotional intensity values. Consequently, incorporating valence-arousal-dominance (VAD) values into the emotional intensity control module offers a promising approach for achieving precise generation of emotional intensity pseudo-labels, along with nuanced control over emotional expression.

Based on above considerations, we propose the Emotional Intensity-aware Network (EINet) that leverages controllable emotional intensity to enhance naturalness and diversity of converted audios, ultimately advancing emotion conversion towards more realistic synthesis. In contrast to solely measuring acoustic features, we focus on the distance between emotional features to construct pseudo-labels for emotional intensity, which effectively addresses the mismatch between emotional intensity modeling and run-time conversion in EVC. To discern nuances in emotional expression at utterance level, the emotion evaluator is employed to anticipate the valence-arousal-dominance (VAD) values behind the speech. Distances between VAD values are further assessed by intensity mapper to obtain pseudo-labels that better align with human-perceived emotional intensity, which highly contributes to enhancing the emodiversity of converted audios. To ensure that speech quality is not compromised by intensity control, the emotion renderer is utilized to integrate linguistic features and manipulated emotional features at frame level. Additionally, we use duration predictor to modify speech duration, adaptively forecasting rhythmic alterations based on emotional intensity values.

Speech samples are available at https://jeremychee4.github.io/EINet4EVC/.

10.21437/Interspeech.2024-1941

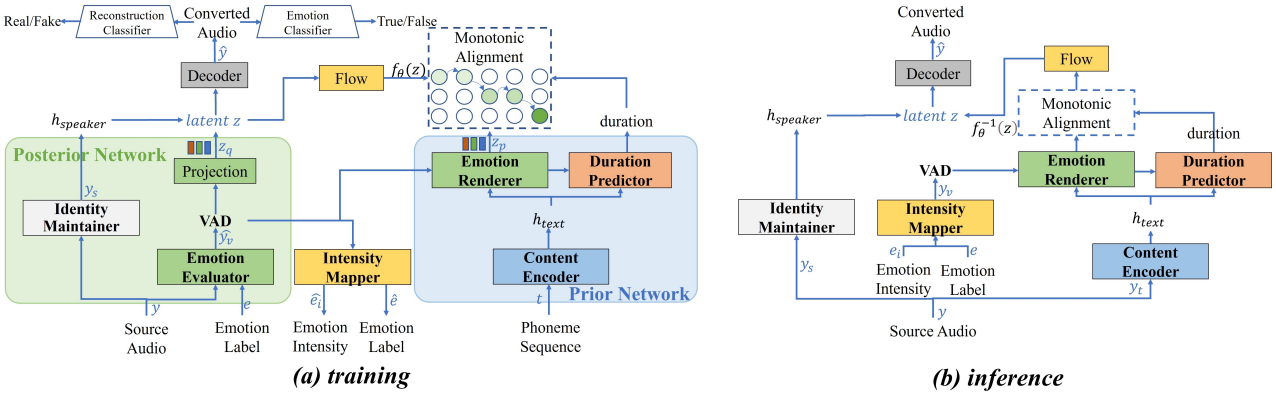

Figure 1: *Diagram of our proposed EINet, depicting the training procedure(a) and inference procedure(b).*

2. Proposed Method

As depicted in Figure 1, EINet is constructed based on conditional variational autoencoder (CVAE), consisting of a posterior network, an intensity mapper, a prior network, and a decoder.

The posterior network (PosNet) captures the inherent emotional states, i.e., valence-arousal-dominance (VAD) values, denoted as y_v, from the source audio y given specific emotion category e, serving as a condition factor for generating posterior distribution $q(z_q \mid c_q)$. Besides, speaker characteristic $h_{speaker}$ is extracted to alleviate the issue of identity loss.

$$z_q = PosNet(c_q) \sim q(z_q \mid c_q) \tag{1}$$

where c_q including source audio y and emotion category e.

The intensity mapper (IM) constructs intensity pseudo-labels $\hat{e}_i$ based on inherent emotional states y_v during training, and generates corresponding VAD values $\hat{y}_v$ given target emotion category e with specified intensity e_i during inference.

$$\begin{cases} \hat{e}, \hat{e}_i = IM(y_v) \\ \hat{y}_v = IM^{-1}(e, e_i) \end{cases} \tag{2}$$

The prior network (PriorNet) predicts prior distribution $p(z_p \mid c_p)$ based on linguistic content y_t and VAD values y_v containing intensity information.

$$z_p = PriorNet(c_p) \sim p(z_p \mid c_p) \tag{3}$$

where c_p including linguistic content y_t and emotional descriptor y_v.

The decoder reconstructs waveform according to latent representation z, where z is derived from z_q during training and z_p during inference, both reinforced with identity information $h_{speaker}$. Additionally, adversarial learning is employed to enhance naturalness progressively in content and emotion aspects.

$$\hat{y} = Decoder(z) \sim p(y \mid z) \tag{4}$$

2.1. Posterior network

Given c_q including source audio y and its emotion category e, the posterior network provides the posterior distribution $q(z_q \mid c_q)$ for CVAE. The emotion evaluator is utilized to extract VAD values at utterance level, which are further transformed to fine-grained emotional acoustic features (a fine-grained normal distribution with mean μ_θ and variance σ_θ generated by a normalizing flow f_θ). Speaker characteristics $h_{speaker}$ is also extracted by identity maintainer.

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

Emotion evaluator: Since speech emotion is inherently supra-segmental, it is difficult to learn emotional latent representation and quantify emotional states in a proper manner. To tackle this, a specific speech emotion recognition (SER) model [21] based on circumplex theory is introduced to predict the valence-arousal-dominance values $\hat{y}_t$ for each utterance, which to assess the pleasantness, activation, and influentiality of speaker's internal states. Utilizing this prior knowledge, emotionally sensitive acoustic features can be extracted by two 1×1 convolutional layers with a Wavenet residual block, and expand to frame-level by a linear projection layer.

Identity maintainer: Considering that controllable EVC has more manipulation over the synthesis of acoustic features, which makes it highly susceptible to speaker identity loss. Recognizing the importance of fundamental frequency (F_0) with voicing flag (v) in modeling speaker characteristics [22], especially for intonation, we enhance the F0 predictor of [23] by incorporating a 1×1 convolutional layer and a linear layer to address this issue.

$$L_{F_0} = ||\log F_0 - \log \hat{F}_0||_2 + ||v - \hat{v}||_2 \tag{6}$$

2.2. Intensity mapper

Different from solely quantifying differences between acoustic features using distance measurement methods, we utilize the intensity mapper to implicitly generate a distribution of intensity pseudo-labels based on emotion category e and VAD values y_v, facilitating supervised training. In order to establish a mapping relationship between intensity distribution and VAD values, it is constructed based on reversible normalizing flow [24].

Intensity label construction: During training, the intensity mapper utilizes VAD values extracted by the emotion evaluator to calculate pseudo-labels $\hat{e}_i \in (0, 1)$ for each sample and predict the emotional category $\hat{e}$.

$$p(\hat{e}, e_i \mid y_v) = N(f_\theta(\hat{e}), f_\theta(e_i); \mu_\theta(y_v); \sigma_\theta(y_v)) \tag{7}$$

To ensure precise mapping and fine-grained control, cross-entropy loss and feature mapping loss are both used to evaluate the accuracy of predictions at both categorical and feature levels. Furthermore, we introduce two coefficients to balance these losses during the early-to-mid and mid-to-late stages.

$$\begin{aligned} L_{IM} = \gamma \mathbb{E}_y \left[-\sum_{k=1}^{K} p_k \log(q_k) \right] \\ + \beta \mathbb{E}_{(y, z_{\mathrm{IM}})} \left[\sum_{l=1}^{L} \frac{1}{N_l} ||D^l(y) - D^l(G(z_{\mathrm{IM}}))||_1 \right] \end{aligned} \tag{8}$$

where K represents the total number of emotional categories, p indicates the emotion label, q represents the classification distribution, L denotes the total number of layers of discriminator, D^l indicates the l-th layer feature map of the discriminator, with N_l denoting the number of features.

Emotional intensity control: During inference, by specifying target emotion category e and intensity value $e_i \in (0, 1)$, the intensity mapper anticipates VAD values $\hat{y}_v$, enabling the direct modulation of emotional expression without the need for any reference audio.

$$\hat{y}_v = f_\theta^{-1}(e, e_i) \tag{9}$$

2.3. Prior network

The prior network provides prior distribution $p(z_p \mid c_p)$ for CVAE based on linguistic content y_t and emotion descriptors y_v. The content encoder takes phoneme sequences as input to extract detailed linguistic feature h_{text} at first. To attain accurate control over emotions, the emotion renderer generates frame-level acoustic features based on VAD values. The duration predictor incorporates emotional and textual features to analyze the correlation among emotional intensity, linguistic sequence, and speech duration, which allows for the prediction of varying durations based on emotional intensity, ultimately enriching the overall emotional diversity.

$$p(z_p \mid c_p) = N\left(f_\theta(z_p); \mu_\theta(c_p); \sigma_\theta(c_p)\right) \left|\det \frac{\partial f_\theta(z_p)}{\partial z_p}\right| \tag{10}$$

Content encoder: To avoid mispronunciations as well as skipping-words that significantly influence human perception, the content encoder plays a crucial role in extracting linguistic features from phoneme sequences, which ensures the preservation of textual content particularly in emotion conversion with intensity control. It comprises a fully connected layer, a Feed-Forward Transformer (FFT) block with a linear projection layer.

Emotion renderer: In order to seamlessly integrate emotional states with linguistic content, the emotion renderer expands generalized VAD values to nuanced emotional acoustic features. It involves a 1×1 dialted convolution layers, a Wavenet residual blocks, and a linear projection layer.

Duration predictor: Considering that diverse emotional intensities can result in distinct voicing durations and pause locations, even when the textual content is the same, we integrate emotional feature and linguistic feature into the duration predictor, aiming to calculate the logarithm of duration at phoneme level, thereby substantially improving the rhythmic modeling capacity of emotional speech. It is constructed using five 1×1 convolution layers, two 1×1 dialted convolution layers and a linear projection layer.

2.4. Final loss

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

$$L = L_{cls} + L_{fm} + L_{adv}(G) + L_{F_0} + L_{dur} + L_{IM} \tag{11}$$

$$L(D) = L_{adv}(D) \tag{12}$$

where L_{cls} minimizes the L1 distance between generated and target spectrogram, L_{fm} minimizes the L1 distance between feature maps extracted from intermediate layers in each discriminator for a better training stability, $L_{adv}(G)$ and $L_{adv}(D)$ represent the adversarial loss for the Generator and Discriminator respectively, L_{dur} minimizes the L2 distance between predicted duration and ground truth which is obtained through estimated alignment.

3. Experiments

3.1. Experimental setup

Dataset. We conduct emotion conversion using a Mandarin corpus within the Emotional Speech Dataset (ESD) [14] from neutral to angry, happy, sad, and surprise, denoted as *Neu-Ang*, *Neu-Hap*, *Neu-Sad*, *Neu-Sur* respectively. The average durations for utterances in each emotional category are 3.23s, 2.68s, 2.84s, 4.04s, and 3.32s, respectively.

Data preparation. For each conversion pair, the corpus is partitioned into a training set (300 samples, approximately 16 minutes), a validation set (30 samples), and a test set (20 samples). In our experiments, we employ speech data represented by an 80-dimensional Mel-spectrogram extracted from audio recorded at a sampling rate of 16kHz.

Implementation details. Our proposed architecture is built upon VITS [25], utilizing the AdamW optimizer with an initial learning rate of 2e-4, and a learning rate decay of 0.999875. Dropout probability is set at 0.1. The γ coefficient starts at 1.00 and decreases by 0.01 every 5 epochs until it reaches 0.30. The β coefficient is defined as $1 - \gamma$.

Models for comparison. We train the following models to assess the effectiveness of proposed method.

- Seq2seq-EVC [12] (*baseline*): a seq2seq-based EVC model supports basic emotion conversion but lacks controllability.
- Emovox [11] (*baseline*): a seq2seq-based EVC model using RAR to calculate the distance between acoustic features, intensity pseudo-labels are obtained to facilitate control.
- VITS-EVC [25] (*baseline*): a EVC model constructed by original VITS, only supports basic emotion conversion.
- EINet (*proposed*): the proposed model utilizing intensity mapper to compute the distance among VAD values, emotional intensity pseudo-labels are obtained to support control.

3.2. Model performance

As depicted in Table 1, we calculate metrics including mel-cepstral distortion (MCD), root mean squared error of log F_0 (RMSE_{F_0}), average differences of duration (DDUR), and classification accuracy from an external pretrained SER model [27] (ACC_{cls}) for objective evaluation. In terms of subjective evaluation, mean opinion score (MOS) test is conducted to assess naturalness and emotion similarity of converted audios by 25 participants, each participant assessing 125 utterances in total.

From above-mentioned indicators, it is obvious that the proposed EINet demonstrates competitive performance in both objective and subjective evaluations. Notably, in comparison to Seq2seq-EVC, Emovox shows minimal improvement in most metrics, particularly with a reduction of 0.11 in naturalness. This implies that relying solely on direct measurement of distances among acoustic feature for intensity pseudo-labels might neglect inherent emotional variations in a speaker, potentially leading to constrained vocal expression during inference. In contrast, EINet achieves more realistic intonation and rhythm variations by utilizing VAD values as guidance, resulting in a reduction of RMSE_{F_0} and DDUR by 4.7 and 0.06, respectively, compared to baseline VITS-EVC. Additionally, there is a significant improvement in naturalness and similarity, which suggests that intensity control module should not compromise basic EVC model when the mismatch between emotion control and speech synthesis is appropriately mitigated. Instead, such controllability has the potential to enhance emotion conversion by capturing more refined emotional cues.

Table 1: *Quantitative comparisons of converted speech with previous methods. The * denotes methods pretrained on VCTK corpus [26].*

EVC Model	Source of Intensity Pseudo-Label	Objective Evaluation				Subjective Evaluation	
		MCD ↓	RMSE$_{F_0}$ ↓	DDUR ↓	ACC$_{cls}$ ↑	Naturalness ↑	Similarity ↑
Seq2seq-EVC*	None	4.22	45.88	0.39	98.85%	4.04±0.16	67.97%
Emovox*	Acoustic features	4.23	47.13	0.36	98.79%	3.93±0.19	66.40%
VITS-EVC	None	4.12	42.92	0.27	99.12%	4.14±0.08	70.07%
EINet (proposed)	VAD values	**4.06**	**38.28**	**0.21**	**99.48%**	**4.38±0.05**	**75.18%**

Table 2: *Ablation study of proposed method.*

EVC Model	RMSE$_{F_0}$ ↓	DDUR ↓	Naturalness ↑
EINet (proposed)	**38.28**	**0.21**	**4.38±0.05**
w/o Identity Maintainer	42.65	0.23	4.29±0.17
w/o Emotion Renderer	41.42	0.29	4.18±0.12
w/o Duration Predictor	46.94	0.38	4.07±0.10

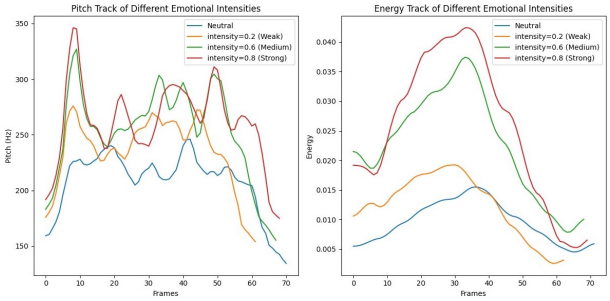

Figure 2: *Pitch and energy tracks of a testing clip.*

3.3. Ablation study

We further conduct an ablation study to validate different contributions. We remove identity maintainer, emotion renderer, and duration predictor in turn and let participants evaluate naturalness of converted audios. From Table 2, we can see that all scores including RMSE$_{F_0}$, DDUR, and naturalness are degraded with the removal of different components. When remove identity maintainer, a significant increase in RMSE$_{F_0}$ is observed. It is attributed that speaker characteristics are not constrained by L_{F_0} in posterior network, which results in unnatural variations in synthesized intonation. To further show the contribution of emotion renderer, we replace it with a simple concatenation, resulting in a slight increase in DDUR, which due to the absence of feature fusion between linguistic content and emotional information before monotonic alignment, thereby weakening prior network's modeling of rhythmic changes. Additionally, the removal of duration predictor leads to a direct impact on all metrics, highlighting the importance of EINet's ability to dynamically adjust speech duration based on the target emotion category and controllable emotional intensity.

3.4. Controllability of emotional intensity

To showcase the controllability of emotional intensity, we visualize pitch and energy tracks of voicing parts in testing clips (from neutral to happy), as exemplified in Figure 2. It can be observed that as emotional intensity increases, i.e., the induction of emotional states progresses from weak to strong, there is a concurrent broadening of pitch fluctuation and an elevation in peak energy. Furthermore, Figure 3 presents synthesized Mel-spectrograms with F0 contours, demonstrating that with an increase in emotional intensity, the acoustic variation becomes more pronounced, coupled with more short pauses. This implies that EINet can adaptively convey intrinsic emotional states based on controllable emotional intensity, achieving optimal outcomes in both intonation and rhythm synthesis.

Table 3: *Diversity evaluation of emotional samples.*

EVC Model	MSD ↑			
	Neu-Ang	Neu-Hap	Neu-Sad	Neu-Sur
Emovox	17.87	16.88	19.86	-
EINet (proposed)	**19.61**	**20.55**	**21.54**	**22.24**

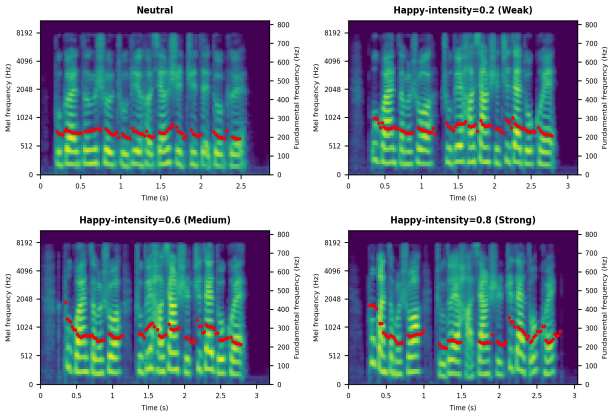

Figure 3: *Mel-spectrograms and F0 contours of converted audios at different emotional intensity.*

3.5. Diversity of emotional samples

The diversity among emotional samples with varying intensity can be quantified through mean squared distance (MSD) metric, which gauges the pairwise distance distribution of converted audio. Table 3 elaborates the MSD values for each emotion conversion. Since Emovox did not conduct experiments on *Neu-Sur* in their paper[11], the presentation of results is consequently absent here.

It is evident that the proposed EINet achieved optimal outcomes for all transformation pairs. This underscores the effectiveness of utilizing VAD values to accurately capture differences in emotional states, offering a valuable solution for addressing the disparity in emotional intensity modeling and runtime conversion. Notably, *Neu-Sad* (long duration, lowest VAD values) and *Neu-Sur* (long duration, moderate valence, high arousal and dominance values) outperform others, indicating that the duration predictor is particularly sensitive to duration and VAD values, when modeling rhythmic variations. Consequently, it generates speech expressions with obvious and natural emotional differences, enhancing overall diversity of converted audios.

4. Conclusion

In this paper, we propose the Emotional Intensity-aware Network (EINet) to achieve realistic emotional voice conversion (EVC) by utilizing controllable emotional intensity. Experimental results on ESD corpus demonstrate its superior performance in enhancing naturalness and diversity of emotional expression, even without explicit emotional intensity annotations. In the future, we will explore the text-based emotion editing for EVC to enhance selectable controllablity of converted audio.

5. Acknowledgements

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, in part by the Jiangsu Frontier Technology Basic Research Project under the Grant BK20192004, and in part by the YESS Program by CAST under the Grant 2023QNRC001.

6. References

[1] Z. Yang, X. Jing, A. Triantafyllopoulos, M. Song, I. Aslan, and B. W. Schuller, "An overview & analysis of sequence-to-sequence emotional voice conversion," *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, pp. 4915–4919, Sep. 2022.

[2] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting anyone's emotion: Towards speaker-independent emotional voice conversion," *Proc. INTERSPEECH 2020 – 21rd Annual Conference of the International Speech Communication Association*, pp. 3416–3420, Oct. 2020.

[3] B. A. Erol, A. Majumdar, P. Benavidez, P. Rad, K.-K. R. Choo, and M. Jamshidi, "Toward artificial emotional intelligence for cooperative social human–machine interaction," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 1, pp. 234–246, 2019.

[4] H. J. Byeon, C. Lee, J. Lee, and U. Oh, ""a voice that suits the situation": Understanding the needs and challenges for supporting end-user voice customization," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–10.

[5] T.-H. Kim, S. Cho, S. Choi, S. Park, and S.-Y. Lee, "Emotional voice conversion using multitask learning with text-to-speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7774–7778.

[6] K. Zhou, B. Sisman, R. Liu, and H. 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.

[7] W. Lu, X. Zhao, N. Guo, Y. Li, J. Wei, J. Tao, and J. Dang, "One-shot emotional voice conversion based on feature separation," *Speech Communication*, vol. 143, pp. 1–9, 2022.

[8] H. Zhu, H. Zhan, H. Cheng, and Y. Wu, "Emotional voice conversion with semi-supervised generative modeling," *Proc. INTERSPEECH 2023 – 24rd Annual Conference of the International Speech Communication Association*, pp. 2278–2282, Aug. 2023.

[9] X. Chen, X. Xu, J. Chen, Z. Zhang, T. Takiguchi, and E. R. Hancock, "Speaker-independent emotional voice conversion via disentangled representations," *IEEE Transactions on Multimedia*, 2022.

[10] T. Qi, W. Zheng, C. Lu, Y. Zong, and H. Lian, "Pavits: Exploring prosody-aware vits for end-to-end emotional voice conversion," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 697–12 701.

[11] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 31–48, 2022.

[12] K. Zhou, B. Sisman, and H. Li, "Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training," *Proc. INTERSPEECH 2021 – 22rd Annual Conference of the International Speech Communication Association*, pp. 811–815, Aug. 2021.

[13] D. Parikh and K. Grauman, "Relative attributes," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 503–510.

[14] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[15] Y. Chen, L. Yang, Q. Chen, J.-H. Lai, and X. Xie, "Attention-based interactive disentangling network for instance-level emotional voice conversion," *Proc. INTERSPEECH 2023 – 24rd Annual Conference of the International Speech Communication Association*, pp. 2068–2072, Aug. 2023.

[16] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3502–3506.

[17] Y. Lei, S. Yang, X. Wang, and L. Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 853–864, 2022.

[18] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *IEEE Transactions on Affective Computing*, 2022.

[19] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[20] R. Reisenzein, "Pleasure-arousal theory and the intensity of emotions." *Journal of personality and social psychology*, vol. 67, no. 3, p. 525, 1994.

[21] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[22] C. Busso, S. Lee, and S. 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.

[23] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, "Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7237–7241.

[24] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3964–3979, 2020.

[25] J. Kim, J. Kong, and J. 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.

[26] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, vol. 6, p. 15, 2017.

[27] Z. Gao, Z. Li, J. Wang, H. Luo, X. Shi, M. Chen, Y. Li, L. Zuo, Z. Du, Z. Xiao, and S. Zhang, "Funasr: A fundamental end-to-end speech recognition toolkit," in *Proc. INTERSPEECH 2023 – 24rd Annual Conference of the International Speech Communication Association*, Dublin, Ireland, Aug. 2023, pp. 1593–1597.