BOOSTING MULTI-SPEAKER EXPRESSIVE SPEECH SYNTHESIS WITH SEMI-SUPERVISED CONTRASTIVE LEARNING

Xinfa Zhu[1]*, Yuke Li*[1]*, Yi Lei*[1]*, Ning Jiang*[2]*, Guoqing Zhao*[2]*, Lei Xie*[1*]

[1]Audio, Speech and Language Processing Group (ASLP@NPU)
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Mashang Consumer Finance Co., Ltd.

ABSTRACT

This paper aims to build a multi-speaker expressive TTS system, synthesizing a target speaker's speech with multiple styles and emotions. To this end, we propose a novel *contrastive learning*-based TTS approach to transfer style and emotion across speakers. Specifically, contrastive learning from different levels, i.e. *utterance* and *category* level, is leveraged to extract the disentangled style, emotion, and speaker representations from speech for style and emotion transfer. Furthermore, a semi-supervised training strategy is introduced to improve the data utilization efficiency by involving multi-domain data, including style-labeled data, emotion-labeled data, and abundant unlabeled data. To achieve expressive speech with diverse styles and emotions for a target speaker, the learned disentangled representations are integrated into an improved VITS model. Experiments on multi-domain data demonstrate the effectiveness of the proposed method.

Index Terms— expressive speech synthesis, style transfer, emotion transfer, contrastive learning, semi-supervised

1. INTRODUCTION

In recent years, neural text-to-speech (TTS) synthesis has rapidly progressed in speech quality and naturalness [1, 2, 3]. With the wide applications of TTS, there have been increasing demands for expressive speech synthesis systems to provide more human-like speech in diverse scenarios. Transfer learning has been the favored method for expressive speech synthesis, which aims to transfer expressiveness from speech recorded by other speakers to the target speaker [4, 5].

In previous works of expressive speech synthesis, speech expressiveness usually refers to specific speaking styles or emotional expressions. These works usually leverage a reference encoder to extract style or emotional expression from reference speech [6, 7]. The critical factor for style or emotion transfer is to decouple the speaker timbre and expressive aspects from speech, as their entanglement usually leads to low speaker timbre similarity or expressiveness [8] for the generated speech. There are many approaches to achieve disentanglement, such as domain adversarial training (DAT) [9], mutual information (MI) [10], and information perturbation [11]. Although researchers utilizing the above approaches have achieved good performance on style or emotion transfer, they have unclear definitions of style and emotion and sometimes consider emotion as a type of speaking style. This indiscriminate treatment of style and emotion restricts them from being extended to real scenarios requiring the combination of different emotions and styles.

Therefore, this paper considers transfer style and emotion from different reference speech simultaneously following the prior investigation [12], which has pointed out that *speaking style* is a general distinctive style of speech in different usage scenarios, such as news reading, storytelling, poetry recitation, and spontaneous conversation. By contrast, *emotion* mainly reflects the mood state of the speaker, related to attitudes and intentions, such as happy, angry, and sad. Building a system that focuses on both speaking style and emotion transfer in multi-speaker expressive speech synthesis usually requires multi-domain datasets containing diverse styles, emotions, and speakers. Meanwhile, speaking style, emotion, and speaker timbre are highly entangled as they all affect the prosody patterns of speech. Even in a reference-based model, the linguistic content of reference speech and texts is consistent during training, which causes entanglement of linguistic content and performance degradation in inference [13], further increasing the difficulty of disentanglement. To solve these problems, this previous study [12] designed a two-stage framework to conduct the disentanglement with neural bottleneck (BN) features as the intermediate representation. Specifically, it proposed a sophisticated disentanglement mechanism by extracting a style/emotion/speaker representation for each category. The disentangled style/emotion/speaker attribute is treated as temporal irrelevant since it is aggregated to a fixed-length vector for the same category. With this assumption, different segments of the same utterance (utterance level) could be involved in extracting the disentangled representations, but this has not been explored in previous works. Moreover, this two-stage framework has many subsystems in which the error accumulation of BN prediction results in the degradation of synthetic speech naturalness.

Contrastive learning is a method to learn the desired features of the data via constructing positive and negative samples. Recently, approaches based on contrastive learning have shown remarkable performance in many fields, such as computer vision (SIMCLR [14], CLIP [15]), reinforcement learning (CURL [16]), and speech processing (MULAN [17], CLAPSpeech [18]). Besides, contrastive learning has shown superiority in speech emotion recognition [19, 20], achieving state-of-the-art performance. By respectively minimizing and maximizing the distances of positive and negative samples [21], contrastive learning has the advantage of extracting latent representations for specific attributes. Besides, through constructing sample pairs from multiple levels, contrastive learning exhibits the potential for performance improvements [22].

With the above considerations, this paper investigates the effectiveness of contrastive learning in the challenging expressive TTS task that enables multi-speaker, multi-style, and multi-emotion speech synthesis. Specifically, this paper proposes a novel contrastive learning-based TTS approach to transfer style and emo-

* Corresponding author.

tion across speakers by extracting the desired attribute representations (i.e., style, emotion, and speaker in this paper). First, we design a Speech Representation Learning (SRL) module to extract style/emotion/speaker-only-related vectors by conducting contrastive learning from both utterance level and category level. Second, we introduce a semi-supervised training strategy to the SRL module, which can effectively leverage multi-domain speech data, including style-labeled, emotion-labeled, and unlabeled data. With this strategy, the SRL can be trained on abundant speech corpus and provide robust style/emotion/speaker representations for TTS. Furthermore, we integrate the learned representations into an improved VITS [3] model and conduct experiments on a multi-domain dataset. Experimental results show that our proposed framework can synthesize diverse stylistic and emotional speech for a target speaker who does not have the target style or emotion in the training data. We suggest readers listen to our online demos [1].

2. PROPOSED APPROACH

The proposed approach comprises an SRL module and a VITS model. Based on contrastive learning, the SRL module aims to extract disentangled style, emotion, and speaker representation from speech. The VITS model synthesizes speech conditioned on the extracted representation.

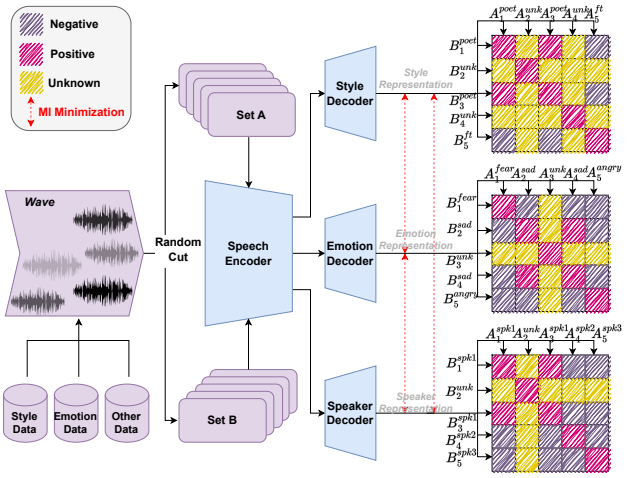

Fig. 1. The architecture of speech representation learning module.

2.1. Semi-supervised positive and negative sample construction

The key to contrastive learning lies in constructing positive and negative samples. Typically, Positive samples are constructed through data augmentation or data from the same category, while negative samples consist of data from different categories [21, 14]. In our approach, the objective is to learn style, emotion, and speaker representations simultaneously. Therefore, multi-domain training data including labeled and unlabeled speech data is available to obtain these disentangled representations. With the diverse training data, we randomly split speech into speech slices and propose a positive and negative sample construction strategy at different levels.

- **Utterance level**: The style, emotion, and speaker timbre are regarded as global attributes, which change slowly alongside the time axis within each single utterance, so different slices of the same utterance can be treated as positive examples of each other.
- **Category level**: Speech samples from the same category exhibit highly correlated style, emotional expression, or speaker timbre. Therefore, speech slices of the same category are positive examples of each other, while speech slices of different categories are negative examples pairs.

With the above utterance-level and category-level sample construction method, we further introduce a semi-supervised training strategy to the contrastive learning, which enables the SRL module to leverage abundant speech data and improve the robustness of the SRL module. Specifically, unlabeled speech data can be used to construct positive samples at the utterance level, while the category-level sample pairs of unlabeled data are not involved during training due to the undefined relationships between the sample pairs. Moreover, randomly Selecting negative samples for unlabeled data will cause adverse effects [23]. Labeled speech data can be leveraged to construct sample pairs at both utterance level and category level.

2.2. Contrastive learning module

As shown in Figure 1, the SRL module consists of a speech encoder, a style decoder, an emotion decoder, and a speaker decoder. The speech encoder comprises a Hubert model and transformer blocks. The Hubert model is to extract features from speech for its exceptional performance across diverse downstream tasks [24]. Transformer blocks encode the Hubert features into three hidden features. These hidden features are then fed to the style, emotion, and speaker decoders to produce global style, emotion, and speaker representations, respectively. We normalize all style, emotion, and speaker representations to a hypersphere by *l2-normalizing*. This normalization eliminates information related to the magnitude and reserves information related to the angular [25], effectively improving the supervision of cosine similarity.

During training, given K speech waveforms, we randomly cut two speech slices from each speech waveform, forming two sets of speech slices, Set A and Set B. We calculate a $K \times K$ cosine similarity matrix $\hat{M}$ between the representation of Set A and Set B, where the value at the position of the i^{th} row and j^{th} column indicates the cosine similarity between the representation of the i^{th} speech slice in Set A and j^{th} speech slice in Set B. The ground truth matrix M consists of -1, 0, and 1 values, where 1 represents positive, 0 represents negative, and -1 represents unknown. The SRL module generates $\hat{M}_{style}$, $\hat{M}_{emotion}$, and $\hat{M}_{speaker}$, and calculate the loss with corresponding the ground truth matrix M_{style}, $M_{emotion}$, and $M_{speaker}$, respectively.

We calculate the contrastive learning loss $\mathcal{L}_{con}$ between $\hat{y}$ in the cosine similarity matrix $\hat{M}$ and y ground-truth matrix M. $\mathcal{L}_{con}$ takes Cross-entropy as the loss function as follows:

$$L_{con}(y,\hat{y}) = \begin{cases} -\log(\hat{y}) & \text{if } y = \text{positive}, \\ -\log(1-\hat{y}) & \text{if } y = \text{negative}, \\ 0 & \text{if } y = \text{unknown}. \end{cases} \quad (1)$$

Moreover, we further disentangle style, emotion, and speaker representation through mutual information (MI) minimization. Given the random variables **u** and **v**, the MI is Kullback-Leibler (KL) divergence between their joint and marginal distributions as $\mathrm{I}(u,v) = D_{KL}(P(u,v);P(u)P(v))$. We adopt vCLUB [26] to compute the

[1] Demo:https://zxf-icpc.github.io/MSES/

upper bound of MI and calculate the MI loss $\mathcal{L}_{MI}$ as:

$$\begin{aligned}\mathcal{L}_{\text{MI}} =& \text{I}(style, emotion) + \text{I}(emotion, speaker) \\ &+ \text{I}(speaker, style)\end{aligned} \tag{2}$$

Therefore, the final training objective of the SRL module $\mathcal{L}_{\text{srl}}$ is as follows:

$$\mathcal{L}_{\text{srl}} = \mathcal{L}_{con} + \mathcal{L}_{\text{MI}} \tag{3}$$

2.3. Expressive VITS

After training the SRL module, a VITS model is trained on conditioning of the extracted style, emotion, and speaker representations by contrastive learning. As shown in Figure 2(a), we use VITS-CLONE [27] as the backbone for its excellent performance on expressive speech synthesis. Specifically, to improve the control ability, we replace the stochastic duration predictor and Monotonic Alignment Search (MAS) module in VITS with the duration predictor and length regulator in FastSpeech2 [2]. Moreover, we use a flow-based prosody adaptor, as shown in Figure 2(b), to capture fine-grained prosody variation of speech and improve the expressiveness of synthetic speech. The prosody adaptor encodes phoneme level $Z_{prosody}$ from the reference mel-spectrogram, which is added with the text encoder output H_{in} to obtain H_{out}.

As shown in Figure 2(a), we add the text encoder output and distribution decoder input with style and emotion representations to control style and emotional expression. Speaker representations are conditioned to the flow, posterior encoder, and decoder. The training objective is the same as that of CLONE. In inference, the SRL module extracts style, emotion, and speaker representation from reference speech, which are then sent to the VITS model with text to synthesize target speech.

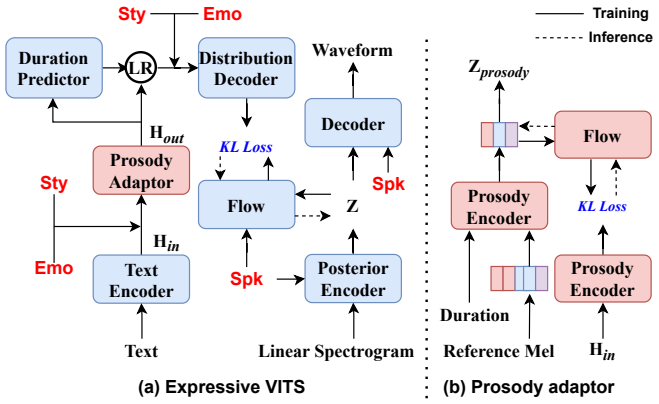

Fig. 2. The architecture of multi-speaker expressive VITS.

3. EXPERIMENTAL SETUPS

3.1. Datasets

There are five corpora involved in the experiments. 1) **CN30S3** contains 18.5 hours of Chinese speech from 30 speakers, where each speaker has one to three styles, including poetry recitation, fairy tales, and storytelling - novels. 2) **CN3E6** contains 21.1 hours of Chinese speech from 30 speakers, and each speaker in this dataset has six emotions: anger, fear, happiness, sadness, surprise, and neutral. 3) **CN5U** has 5.8 hours of Mandarin speech from 5 speakers. 4) **EN5U** 31.3 hours of English speech from 5 speakers. 5) **MIXU** has 900 hours of Chinese and English speech collected from internal resources without annotations and transcripts.

For all recordings, we down-sample them into 24k Hz and set the frame and hop size to 1200 and 300, respectively. We cut recordings into 3 seconds of speech slices to construct the sample pairs of contrastive learning. For that speech of less than 3 seconds, we repeat them in the time axis until they are more than 3 seconds. We use a Bilingual TTS front end to encode Chinese text input into phonemes, tones, word boundaries, and prosodic boundaries while decoding English text input into phonemes. We leverage Pinyin and CMU-Dict as the phoneme set. The phoneme duration is obtained through an HMM-based force alignment model [28].

3.2. Model configuration

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

- **TSEW** [12]: A two-stage framework with a **T**ext-to-**S**tyle-and-**E**motion module and a style-and-emotion-to-**W**ave module. The former module predicts neural bottleneck features with style and emotion, while the latter predicts waveforms from bottleneck features conditioned on the speaker and emotion embedding.
- **SCVITS** [19]: VITS model with a **S**upervised-**C**ontrastive learning module. Specifically, we replace Wav2Vec2.0 with Hubert for fair comparison and train three models to learn emotion, style, and speaker representations on the corresponding labeled dataset, respectively. We train the VITS model on these learned representations with the same configuration as the proposed approach.
- **Proposed**: The proposed framework with an SRL module and VITS model.

In our implementation, we use the Chinese-Hubert-Large [2] to extract features from layer 6 to layer 18. We find that the Hubert features from a single layer, such as layer 6, perform worse than those from multiple layers in our experiments. Transformer blocks consist of 3 layers, 2 attention heads, an embedding dimension of 256, a feed-forward layer dimension of 1024, and a dropout of 0.2. The structure of style, emotion, and speaker decoders is the same and follows the structure of the reference encoder in the prior work [12]. The mutual information estimator also follows the settings of [12]. The backbone of expressive VITS keeps the same configuration as CLONE.

We train the SRL module with a batch size of 96 and the expressive VITS model with a batch size of 48. To balance the labeled and unlabeled data during the training of the SRL module, the batch is divided into four equal parts. One-fourth of the batch comes from style-labeled data, another-fourth comes from emotion-labeled data, and the third and fourth parts come from speaker-labeled and unlabeled data, respectively.

3.3. Evaluation metrics

For monolingual TTS evaluation, given 20 reserved transcripts for each style, we generate samples respectively for each emotion category, resulting in 360 listening samples per person (20 texts $\times$ 3 styles $\times$ 6 emotions). We randomly select two speakers from CN5U

[2]HuBERT:https://github.com/TencentGameMate/chinese_speech_pretrain

Table 1. Results of monolingual subjective evaluation with 95% confidence interval and objective evaluation.

Model	Naturalness↑	Emotion Similarity↑	Speaker Similarity↑	Style Similarity↑	CER(%)↓	SCS↑
TSEW	3.94 ± 0.10	3.91 ± 0.11	3.88 ± 0.07	3.85 ± 0.11	6.2	0.866
SCVITS	3.97 ± 0.08	3.84 ± 0.10	3.75 ± 0.11	3.69 ± 0.11	4.8	0.884
Proposed	$\mathbf{4.09 \pm 0.08}$	$\mathbf{3.96 \pm 0.11}$	$\mathbf{3.97 \pm 0.07}$	$\mathbf{4.03 \pm 0.11}$	**3.0**	**0.909**
- MI	4.01 ± 0.09	3.88 ± 0.10	3.90 ± 0.09	3.88 ± 0.11	4.5	0.893

as target speakers. For multilingual TTS evaluation, we add 20 English transcripts and generate samples for each emotion category, resulting in 120 listening samples. We randomly select a speaker from each of CN5U and EN5U as target speakers.

For subjective evaluation, We conduct Mean Opinion Score (MOS) experiments to evaluate speech naturalness and Similarity Mean Opinion Scores (SMOS) to evaluate emotion similarity, speaker similarity, and style similarity, respectively. Twenty volunteers with basic bilingual skills take part in the assessment. During the evaluation, participants are told to focus on specific aspects while ignoring others. For objective evaluation, we use an ECAPA-TDNN [29] model trained on 3,300 hours of Mandarin speech and 2,700 hours of English speech from 18,083 speakers to measure speaker cosine similarity (SCS). Moreover, we use an open-source U2++ conformer model provided by the WeNet community [30] to evaluate character error rate (CER), and word error rate (WER). The U2++ conformer model is trained on 10,000 hours of open-source Gigaspeech English data and WeNet Mandarin data, respectively.

4. EXPERIMENTAL RESULTS

We first evaluate the performance of the proposed approach on monolingual corpora (CN30S3, CN3E6, and CN5U). Then, we conduct experiments on multilingual corpora (CN30S3, CN3E6, CN5U and EN5U), examining the effectiveness of the proposed approach in the cross-lingual transfer setting.

4.1. Monolingual subjective evaluation

As shown in Table 1, the proposed approach outperforms the compared models in terms of naturalness, emotion similarity, speaker similarity, and style similarity. TSEW gets the lowest naturalness, which we speculate that the error accumulation of BN prediction leads to this phenomenon. SCIVTS achieve higher naturalness but the lowest emotion, speaker, and style similarity. The style, emotion and speaker representations are extracted from three independent modules, which are probably entangled and lead to low similarity. These results show that our proposed approach obtains well-disentangled emotion, style, and speaker representations. Besides, the end-to-end VITS model avoids the error accumulation of intermediate representations, enabling flexible and natural expressive speech synthesis.

Moreover, to verify the advantages of the proposed method, we remove the mutual information estimator in the SRL module. As shown in Table 1, the model (-MI) still outperforms the compared models, showing the effectiveness of contrastive learning. With the mutual information estimator, the SRL module can better disentangle the style, emotion, and style representation, and the whole model obtains better performance.

4.2. Monolingual objective evaluation

As shown in Table 1, the proposed approach achieves the lowest CER, showing the robustness of the proposed framework. Moreover, the proposed approach obtains the highest SCS, indicating the speaker characteristics are well captured and disentangled. TSEW gets the worst CER, which we conjecture is due to the two-stage framework and the error accumulation of BN prediction. Moreover, SCVITS gets lower speaker cosine similarity, proving that the representation learned by SCVITS is not well-disentangled. Removing MI from the proposed framework leads to a performance decline in all objective evaluations, demonstrating the effectiveness of MI in achieving accurate pronunciation and high speaker cosine similarity.

Furthermore, to verify the effectiveness of the SRL module, we visualize the emotion and style representations through t-SNE [31]. One hundred fifty utterances reserved per emotion and 250 per style are adopted for test. As shown in Figure 3, the style and emotional representations are well clustered by corresponding categories, indicating the effectiveness of the SRL model. Moreover, the style and emotional representations can not be clustered by speaker identities, showing the good disentanglement between speaker and style, speaker and emotion.

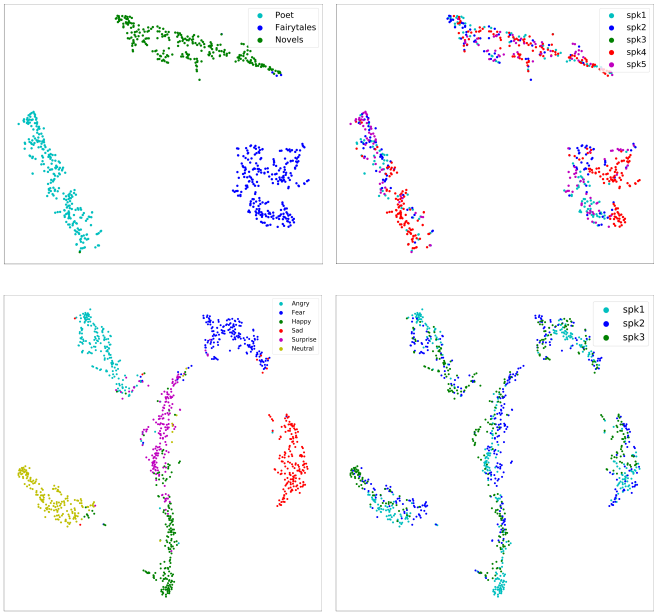

Fig. 3. T-SNE visualization of style representation (above) and emotion representation (below). We color the results with the corresponding category (left) and speaker category (right).

4.3. Multilingual subjective evaluation

The subjective evaluation results of multilingual TTS are shown in Table 2; all models exhibit a performance degradation compared to Table 1, revealing the challenges of cross-lingual expressive speech

Table 2. Results of multilingual subjective evaluation with 95% confidence interval and objective evaluation.

Model	Naturalness↑	Emotion Similarity↑	Speaker Similarity↑	Style Similarity↑	CER(%)↓	WER(%)↓	SCS↑
TSEW	3.76 ± 0.12	3.87 ± 0.10	3.61 ± 0.11	3.77 ± 0.10	6.8	9.7	0.847
SCVITS	3.84 ± 0.08	3.80 ± 0.12	3.70 ± 0.11	3.65 ± 0.12	5.7	4.9	0.838
Proposed	4.01 ± 0.07	3.92 ± 0.10	3.90 ± 0.08	4.01 ± 0.09	**3.9**	**2.7**	0.896
- MI	3.92 ± 0.10	3.84 ± 0.12	3.77 ± 0.09	3.84 ± 0.12	5.2	3.6	0.852
+ MIXU	$\mathbf{4.03 \pm 0.08}$	$\mathbf{3.98 \pm 0.11}$	$\mathbf{3.95 \pm 0.08}$	$\mathbf{4.04 \pm 0.10}$	3.9	2.8	**0.903**

synthesis. However, the proposed approach demonstrates relatively minor degradation in performance during cross-lingual expressive speech synthesis, suggesting its capability to generate fluent and expressive foreign speech for a given target speaker. TSEW gets the lowest naturalness due to the unnatural pronunciation in synthetic English speech, which also affects the listeners' judgment in speaker similarity evaluation. Primarily, BN in TSEW is extracted through a robust TDNN-F model trained with 30k hours of Chinese speech data, resulting in inaccurate pronunciation in English speech synthesized by TSWE. Moreover, SCVITS obtains a serious performance degradation. The supervised contrastive learning module of SCVITS can only be trained on labeled data, which means English speech is unseen during training and causes performance degradation. These results show the effectiveness of multi-level contrastive learning and semi-supervised training strategy in cross-lingual expressive speech synthesis.

Removing the mutual information estimator in the SRL module encounters a slight performance decline, which is consistent with monolingual evaluation. Besides, to evaluate the crucial role of the semi-supervised training strategy, we add the corpus MIXU during the SRL module training. With enlarged training corpora, the overall performance of the proposed system is improved, which means the wealth of variation in abundant unlabeled speech data helps capture more precise style, emotion and speaker characteristics.

4.4. Multilingual objective evaluation

As shown in Table 2, the proposed approach achieves the lowest CER and WER and the highest SCS, indicating the robustness of the proposed approach in cross-lingual expressive speech synthesis. TSEW gets the worst WER as BN is extracted through the TDNN-F model trained on Chinese speech data and the pronunciation is inaccurate. Additionally, SCVITS fails to effectively address the challenge of emotion, style, and speaker entanglement in multilingual settings, yielding low SCS and high CER and WER. Removing MI from the proposed framework leads to a performance decline in all objective evaluations while adding the corpus MIXU improves overall performance. These results confirm the observations from the subjective evaluation.

Moreover, to study the relationship between style, speaker and language, we visualize multilingual emotion and style representations through t-SNE [31]. One thousand utterances reserved per language are adopted for test. As shown in Figure 4, the style representation tends to be language-specific while the emotion representation seems to be language-agnostic. We speculate that basic emotional expressions such as happiness and sadness are available in all languages [32, 33], causing the emotion representation to be language-agnostic. However, different manners of pronunciation in different languages lead to different speaking styles, which results in language-specific style representations [34].

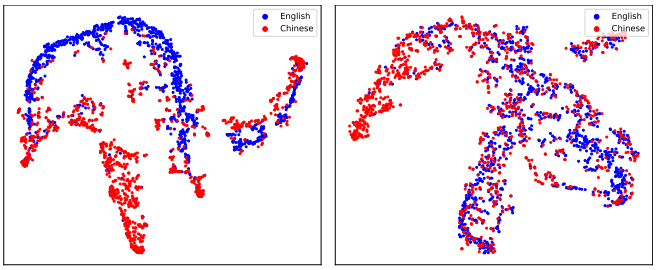

Fig. 4. T-SNE visualization of style representation (left) and emotion representation (right) in multilingual settings.

5. CONCLUSIONS

This paper aims to synthesize speech with the desired style and emotion for target speakers by transferring the style and emotion from reference speech recorded by other speakers. We approach this challenging problem with a novel contrastive learning-based TTS framework. Specifically, this paper proposes a novel speech representation learning module based on contrastive learning, which constructs sample pairs at utterance and category levels and learns disentangled style, emotion, and speaker representations. Besides, we introduce a semi-supervised training strategy to the proposed framework, which leverages multi-domain data and helps learn robust representations. We integrate the learned style, emotion, and speaker representation into an improved VITS model and conduct experiments on monolingual and multilingual datasets. Extensive experimental results demonstrate the proposed framework can synthesize speech with diverse speaking styles and emotions for a target speaker, even if the speaking style or emotion comes from another language.

6. REFERENCES

[1] Xu Tan, Tao Qin, Frank K. Soong, and Tie-Yan Liu, "A survey on neural speech synthesis," 2021, vol. abs/2106.15561.

[2] 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 *Proc. ICLR*. 2021, OpenReview.net.

[3] Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*. 2021, pp. 5530–5540, PMLR.

[4] Alexander Sorin, Slava Shechtman, and Ron Hoory, "Principal style components: Expressive style control and cross-speaker transfer in neural TTS," in *Proc. Interspeech*. 2020, pp. 3411–3415, ISCA.

[5] Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for

emotional speech synthesis," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 853–864, 2022.

[6] Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang, "Meta-stylespeech : Multi-speaker adaptive text-to-speech generation," in *Proc. ICML*. 2021, pp. 7748–7759, PMLR.

[7] Tao Li, Shan Yang, Liumeng Xue, and Lei Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *Proc. ISCSLP*. 2021, pp. 1–5, IEEE.

[8] Tao Li, Xinsheng Wang, Qicong Xie, Zhichao Wang, and Lei Xie, "Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 1448–1460, 2022.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, pp. 59:1–59:35, 2016.

[10] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng, "VQMIVC: vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," in *Proc. Interspeech*. 2021, pp. 1344–1348, ISCA.

[11] Yi Lei, Shan Yang, Xinfa Zhu, Lei Xie, and Dan Su, "Cross-speaker emotion transfer through information perturbation in emotional speech synthesis," 2022, vol. 29, pp. 1948–1952.

[12] Xinfa Zhu, Yi Lei, Kun Song, Yongmao Zhang, Tao Li, and Lei Xie, "Multi-speaker expressive speech synthesis via multiple factors decoupling," in *Proc. ICASSP*, 2023, pp. 1–5.

[13] Yi Meng, Xiang Li, Zhiyong Wu, Tingtian Li, Zixun Sun, Xinyu Xiao, Chi Sun, Hui Zhan, and Helen Meng, "CALM: constrastive cross-modal speaking style modeling for expressive text-to-speech synthesis," in *Proc. Interspeech*. 2022, pp. 5533–5537, ISCA.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*. 2020, vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, PMLR.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. ICML*. 2021, vol. 139, pp. 8748–8763, PMLR.

[16] Michael Laskin, Aravind Srinivas, and Pieter Abbeel, "CURL: contrastive unsupervised representations for reinforcement learning," in *Proc. ICML*. 2020, vol. 119 of *Proceedings of Machine Learning Research*, pp. 5639–5650, PMLR.

[17] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis, "Mulan: A joint embedding of music audio and natural language," in *Proc. ISMIR*, 2022, pp. 559–566.

[18] Zhenhui Ye, Rongjie Huang, Yi Ren, Ziyue Jiang, Jinglin Liu, Jinzheng He, Xiang Yin, and Zhou Zhao, "Clapspeech: Learning prosody from text context with contrastive language-audio pre-training," in *Proc. ACL*, 2023, pp. 9317–9331.

[19] Varun Sai Alaparthi, Tejeswara Reddy Pasam, Deepak Abhiram Inagandla, Jay Prakash, and Pramod Kumar Singh, "Scser: Supervised contrastive learning for speech emotion recognition using transformers," in *Proc. HSI*. 2022, pp. 1–7, IEEE.

[20] Mao Li, Bo Yang, Joshua Levy, Andreas Stolcke, Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannis, Daniel Bone, and Chao Wang, "Contrastive unsupervised learning for speech emotion recognition," in *Proc. ICASSP*. 2021, pp. 6329–6333, IEEE.

[21] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon, "A survey on contrastive self-supervised learning," *CoRR*, vol. abs/2011.00362, 2020.

[22] Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li, "Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition," *CoRR*, vol. abs/2308.04502, 2023.

[23] Jaejin Cho, Jesús Villalba, Laureano Moro-Velázquez, and Najim Dehak, "Non-contrastive self-supervised learning for utterance-level information extraction from speech," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1284–1295, 2022.

[24] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[25] Ji-Hoon Kim, Sang-Hoon Lee, Ji-Hyun Lee, Honggyu Jung, and Seong-Whan Lee, "GC-TTS: few-shot speaker adaptation with geometric constraints," in *Proc. SMC*. 2021, pp. 1172–1177, IEEE.

[26] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin, "CLUB: A contrastive log-ratio upper bound of mutual information," in *Proc. ICML*. 2020, pp. 1779–1788, PMLR.

[27] Zhengxi Liu, Qiao Tian, Chenxu Hu, Xudong Liu, Menglin Wu, Yuping Wang, Hang Zhao, and Yuxuan Wang, "Controllable and lossless non-autoregressive end-to-end text-to-speech," *CoRR*, vol. abs/2207.06088, 2022.

[28] Kåre Sjölander, "An hmm-based system for automatic segmentation and alignment of speech," 2003.

[29] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*. 2020, pp. 3830–3834, ISCA.

[30] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech*. 2021, pp. 4054–4058, ISCA.

[31] Laurens van der Maaten and Geoffrey E. Hinton, "Visualizing data using t-sne," 2008, vol. 9, pp. 2579–2605.

[32] Keshi Dai, Harriet J. Fell, and Joel MacAuslan, "Comparing emotions using acoustics and human perceptual dimensions," in *Proc. CHI*. 2009, pp. 3341–3346, ACM.

[33] Björn W. Schuller, Ronald Müller, Manfred K. Lang, and Gerhard Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Proc. Interspeech*. 2005, pp. 805–808, ISCA.

[34] Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *CoRR*, vol. abs/2303.03926, 2023.