



# Self-supervised Context-aware Style Representation for Expressive Speech Synthesis

Yihan Wu<sup>1,†</sup>, Xi Wang<sup>2</sup>, Shaofei Zhang<sup>2</sup>, Lei He<sup>2</sup>, Ruihua Song<sup>1,\*</sup>, Jian-Yun Nie<sup>3</sup>

<sup>1</sup> Gaoling School of Artificial Intelligence, Renmin University of China, China

<sup>2</sup> Microsoft Azure Speech, China

<sup>3</sup> Université de Montréal, China

yihanwu@ruc.edu.cn, {xwang, shazh, helei}@microsoft.com, nie@iro.umontreal.ca

## Abstract

Expressive speech synthesis, like audiobook synthesis, is still challenging for style representation learning and prediction. Deriving from reference audio or predicting style tags from text requires a huge amount of labeled data, which is costly to acquire and difficult to define and annotate accurately. In this paper, we propose a novel framework for learning style representation from abundant plain text in a self-supervised manner. It leverages an emotion lexicon and uses contrastive learning and deep clustering. We further integrate the style representation as a conditioned embedding in a multi-style Transformer TTS. Comparing with multi-style TTS by predicting style tags trained on the same dataset but with human annotations, our method achieves improved results according to subjective evaluations on both in-domain and out-of-domain test sets in audiobook speech. Moreover, with implicit context-aware style representation, the emotion transition of synthesized audio in a long paragraph appears more natural. The audio samples are available on the demo website. <sup>1</sup>

**Index Terms:** expressive TTS, self-supervised learning, deep clustering, representation learning, contrastive learning

## 1. Introduction

Although TTS models can synthesize clean and high-quality natural speeches, it still suffers from the issue of over-smoothing prosody pattern in some complex scenarios, as in audiobook synthesis. One of the reasons is the difficulty of modeling high-level characteristics such as emotions and context variations, which impact the overall prosody and speaking style. Being different with the low-level acoustic characteristics such as duration, pitch and energy, modeling high-level characteristics is more challenging and crucial in these complex scenarios [1].

There are two general approaches to deal with such tasks: unsupervised joint training and supervised label conditioning. The unsupervised approach models styles based on joint training with both reference audio and text content [1, 2, 3, 4, 5]. By constructing an implicit style representation space in an unsupervised way, it infers a style representation from either the reference audio or the predicted style by the joint training process. However, the joint training framework faces two challenges: 1) content information leaks into style encoder; 2) requiring a large number of audio and content pairs. Many recent studies in this area focus on these two issues [6, 7, 8]. In real applications, the

supervised learning is more widely adopted by leveraging explicit labels as auxiliary information to guide multi-style TTS [9, 10]. It does not require reference audio, but the definition of styles, which could be subjective. Predicting style tags also requires a large amount of annotated data. Moreover, a simple discrete tag cannot fully reflect the nuance in speech styles.

To address these problems, instead of modeling styles through reference audios or explicit tags, we propose a novel framework which learns the style representation from plain text in a self-supervised manner and integrates it into an end-to-end conditioned TTS model. First, we employ contrastive learning to pre-train style embedding by distinguishing between similar and dissimilar utterances. To this end, we create a similar utterance by replacing an emotional word by a similar one, determined using an emotion lexicon. With the emotionally similar utterance as positive sample, all other dissimilar utterances in the randomly sampled minibatch are treated as negatives. Then training samples in style embedding space are clustered by minimizing deep clustering loss, reconstruction loss and contrastive loss together. We learn the style representation from a large amount of unlabeled plain text data and construct a text sentiment embedding space to guide the generation of multi-style expressive audio in speech synthesis. Using it as a pre-training of style information, we can get rid of the dependence of matched audio and content. Our work has three main contributions:

- We propose a novel framework for modeling style representation from unlabeled texts and incorporate it into a style-based TTS model, without reference audio or explicit style labels.
- We propose a novel two-stage style representation learning method combining deep embedded clustering with contrastive learning based on data augmented via an emotion lexicon.
- We demonstrate that with the same labeled text corpus and audiobook corpus, our speech synthesis outperforms the baseline, especially in naturalness of emotion transition in long audio generation.

## 2. Related Work

Our work is related to contrastive learning and deep clustering.

Contrastive learning is one of the most efficient ways to extract useful information from massive unlabeled data. Chen et al. [11] propose a simple contrastive framework called simCLR which learns visual representations by maximizing the similarity between a similar pair. Furthermore, SimSiam [12] explores simple siamese network to learn meaningful representation without negative sample pairs. Besides, contrastive learning is also widely used in NLP and speech areas [13, 14, 15, 16].

<sup>†</sup> Work done during an internship at Microsoft.

\* Corresponding author: songruihua.bloon@outlook.com.

<sup>1</sup> <https://wyh2000.github.io/InterSpeech2022/>

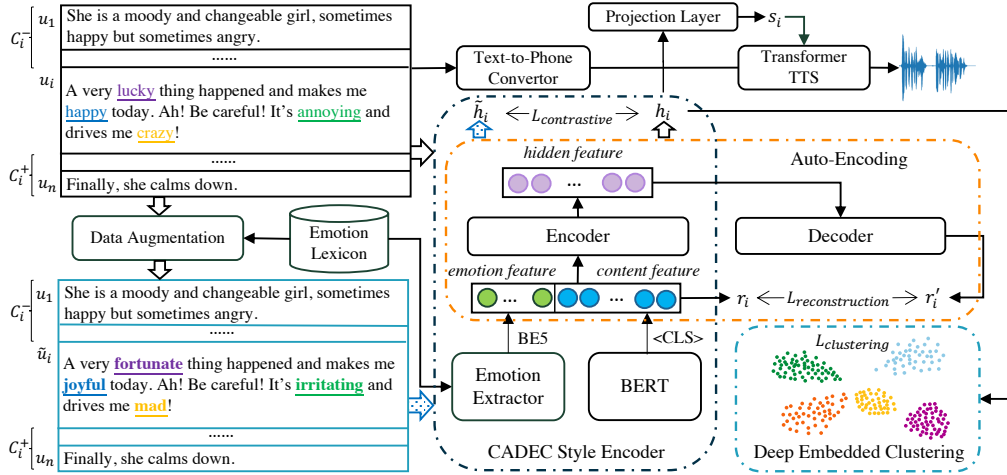


Figure 1: Overview of our proposed framework

Compared to the existing work, we propose a new data augmentation strategy based on emotion strength and apply simCLR to our scenario.

Clustering is an effective method of analyzing data without category annotations. Deep embedded clustering (DEC) [17] maps the observed data to a low-dimensional space and optimizes KL divergence as clustering objective. Improved deep embedded clustering (IDEC) [18] adds reconstruction loss to maintain original data structure. Many extension models have been proposed, leading to high performance in various tasks [19, 20, 21]. Combining with contrastive learning, Supporting Clustering with Contrastive Learning (SCCL) [22] achieves state-of-the-art results on short text clustering task by jointly optimizing clustering loss and contrastive loss. Inspired by these studies, we combine contrastive learning and deep clustering to improve expressive TTS systems for the first time.

### 3. Our Approach

#### 3.1. Problem Formulation and System Overview

Assume a text dataset  $\mathcal{D} = \{U_i\}_{i=1}^D$  where  $U_i = \{C_i^-, u_i, C_i^+\}$  represents text utterance  $u_i$  and its context.  $C_i^- = \{u_{i-m}, \dots, u_{i-1}\}$  is the preceding utterances of  $u_i$  and  $C_i^+ = \{u_{i+1}, \dots, u_{i+m}\}$  is the following utterances of  $u_i$ . Our goal is to learn a style encoding model  $s_i = g(U_i)$  from  $\mathcal{D}$  which can generate a context-aware style representation for utterance  $u_i$ . This style model will be applied to a TTS system to improve the expressiveness of speech synthesis.

Figure 1 shows our proposed framework. First, we construct positive pairs of  $u_i$  and its augmented sample  $\tilde{u}_i$  by replacing the words having the strongest emotion arousal by their synonyms. Based on the data, we design a style encoder and pre-train it via contrastive learning. Second, to optimize the global distribution of our style representation, we further enhance the improved Deep Embedded Clustering method [18] with contrastive learning to train our style encoder further. Through the two stages, we learn  $g(U_i)$  and denote the generated representation as Context-aware Augmented Deep Embedded Clustering (CADEC) style. Finally, we feed it into Transformer TTS as conditioning embedding to generate expressive audio by applying appropriate style to text.

#### 3.2. Stage 1: Contrastive Learning with Data Augmentation

An utterance could be expressed by human in various styles. The appropriate style of utterance  $u_i$  is highly related to context, its semantic content and conveyed emotion. We propose taking  $u_i$  and its context together, i.e.  $U_i$ , as input and combine both content feature and emotion feature to model the best-fit style.

We employ a pretrained BERT [23] as backbone to extract content features, and an extra emotion lexicon [24] to extract emotion features. The emotion lexicon starts from a manually annotated English source emotion lexicon. Combining emotion mapping, machine translation, and embedding-based lexicon expansion, the monolingual lexicons for 91 languages with more than two million entries for each language are created. The lexicon provides word-level emotion features including VAD (valance, arousal, dominance) on 1-to-9 scales and BE5 (joy, anger, sadness, fear, disgust) on 1-to-5 scales. Then, we extract our initial style embedding  $r_i$  by:

$$r_i = b(U_i) \oplus \frac{1}{M} \sum_{j=1}^M e(w_j) \quad (1)$$

where  $\oplus$  denotes a concatenation operator,  $b(U_i)$  is the output [CLS] embedding by inputting  $U_i$  into BERT,  $M$  is the total number of words in  $U_i$  and  $w_j$  is  $j$ -th word in  $U_i$  while  $e(w_j)$  denotes its normalized BE5 feature which is a 5-dimensional vector.

Then we add a fully connected multilayer perceptron (MLP) as encoder to map the initial embedding into hidden features, which are our output style embedding:

$$h_i = MLP(r_i) \quad (2)$$

We propose augmenting data and using contrastive learning to pre-train the parameters of encoder.

To augment  $u_i$  to the utterance  $\tilde{u}_i$  that would have similar speech style, we first split  $u_i$  into shorter segments not longer than a fixed length, e.g., 10 in our experiments. Then we look up the emotion lexicon to get emotion arousal for each word in a segment and select top  $k\%$ , e.g., 20%, to be replaced by their WordNet synonyms [25]. Take the utterance  $u_i$  in Figure 1 as an example. We split it into two segments, and select ‘‘lucky’’ and ‘‘happy’’ in the first segment and ‘‘annoying’’ and ‘‘crazy’’ in the second segment. We then replace them with their synonyms to compose  $\tilde{u}_i$ . The aim of splitting a long sentence into segments

Table 1: *subjective evaluation of proposed model and baseline on TTS-evaluation set*

Metrics		MOS			CMOS		Paragraph CMOS	
Settings		Recording	Baseline	Our Model	Baseline	Our Model	Baseline	Our Model
in-domain		4.35 ± 0.01	4.26 ± 0.07	4.34 ± 0.06	0	<b>+0.22</b>	—	—
out-of-domain	Female	4.37 ± 0.1	4.21 ± 0.07	4.28 ± 0.06	0	<b>+0.03</b>	0	<b>+0.22</b>
	Male	4.35 ± 0.13	4.1 ± 0.1	4.18 ± 0.09	0	<b>+0.05</b>	0	<b>+0.23</b>

is to extract emotional words from different segments, thereby avoiding focusing on the dominant emotional words from some segment only. For example, although “fortunate” has higher arousal than “annoying” in the whole sentence of 20 words, we avoid choosing it for the whole sentence by the segment-based selection. This makes our concentrated emotional words more evenly distributed to ensure the expressiveness of the whole sentence.

As for contrastive learning, from a large training dataset  $\mathcal{D}$ , we randomly sample a minibatch data  $\mathcal{B} = \{U_i\}_{i=1}^N$ , and generate its augmented data  $\tilde{\mathcal{B}} = \{\tilde{U}_i\}_{i=1}^N$ , where  $\tilde{U}_i = \{C_i^-, \tilde{u}_i, C_i^+\}$ .  $U_i$  and  $\tilde{U}_i$  are treated as positive pairs while the other  $N-1$  pairs  $\{\langle U_i, \tilde{U}_k \rangle\}_{i \neq k}$  are all negative examples in one minibatch. To maximize the agreement between texts with similar emotions and disagreement between texts with different emotions, following simCLR [11], we calculate the sample-wise contrastive loss by

$$\ell_c^i = -\log \frac{\exp(\cos(h_i, \tilde{h}_i)/\tau)}{\sum_{k=1}^N \mathbb{1}_{k \neq i} \exp(\cos(h_i, \tilde{h}_k)/\tau)} \quad (3)$$

Here  $\tau$  is the temperature parameter and  $\mathbb{1}_{k \neq i}$  is the indicator function. The contrastive loss for a minibatch is computed by averaging over all instances in  $\mathcal{B}$  and its augmented data  $\tilde{\mathcal{B}}$ :

$$\mathcal{L}_{contrastive} = \frac{1}{N} \sum_i \ell_c^i \quad (4)$$

As the significant overlap of initial representation, this stage proves useful as the start of Stage 2 in our experiments.

### 3.3. Stage 2: Deep Embedded Clustering with Autoencoder

To optimize the global distribution of style representations, we apply deep embedded clustering with autoencoder to train the CADEC style encoder further. The number of clusters  $K$  is a prior and each cluster is represented by its centroid  $\mu_k$ . Clustering loss is defined as

$$\mathcal{L}_{clustering} = KL(P||Q) = \sum_i \sum_k p_{ik} \log \frac{p_{ik}}{q_{ik}} \quad (5)$$

where  $P$  is the target distribution of  $Q$ . Following [26], we apply the Student’s  $t$ -distribution to compute the probability of assigning  $h_i$  to the  $k^{th}$  cluster  $q_{ik}$ .

$$q_{ik} = \frac{(1 + \|h_i - \mu_k\|_2^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k'=1}^K (1 + \|h_i - \mu_{k'}\|_2^2/\alpha)^{-\frac{\alpha+1}{2}}} \quad (6)$$

where  $\alpha$  denotes the degree of freedom of the Student’s  $t$ -distribution. In this work, we set  $\alpha = 1$ . The target distribution  $p_{ik}$  is

$$p_{ik} = \frac{q_{ik}^2 / \sum_i q_{ik}}{\sum_{k'} (q_{ik'}^2 / \sum_i q_{ik'})} \quad (7)$$

As this kind of clustering distorts the original space of representation and weakens the representation ability of implicit

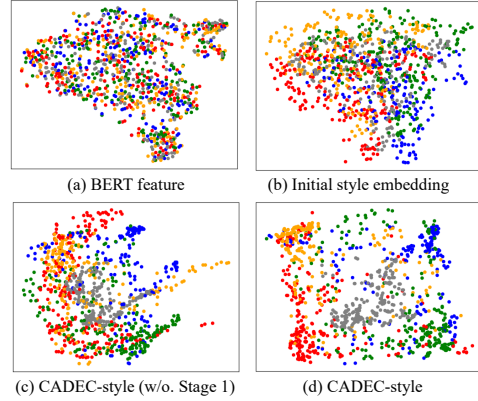


Figure 2: *Visualization of the embedding space learned from different models. Each color indicates a ground truth emotion category. (a) represents the original BERT [CLS] embedding without fine-tuning. (b) represents initial style embedding which combines BERT embedding with BE5. (c) represents style embedding learned from our proposed model without Stage 1. (d) represents our learned CADEC style embedding.*

feature [18], we add an autoencoder structure with reconstruction loss as in Equation 8, in order to preserve local structure of feature space and avoid corruption:

$$\mathcal{L}_{reconstruction} = \sum_{i=1}^N \|r_i - r'_i\|_2^2 \quad (8)$$

Therefore, the objective of Stage 2 is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{contrastive} + \beta \mathcal{L}_{clustering} + \gamma \mathcal{L}_{reconstruction} \quad (9)$$

We tune all trainable parameters to optimize  $\mathcal{L}_{total}$  and obtain the final style encoder of our CADEC style.

### 3.4. TTS Stage: Transformer TTS with Style Representation

Transformer TTS [27] leverages Transformer-based encoder-attention-decoder architecture. It performs well in synthesizing high-quality audio in fast speed. We extend Transformer TTS architecture by conditioning on CADEC style embedding generated from our proposed style encoder, with phoneme sequence and its style representation as inputs, as shown in Figure 1. Due to space limitation, we refer the reader to the paper [27] for more details about conditioned Transformer TTS.

## 4. Experiments

### 4.1. Setup

To train and evaluate CADEC style encoder and TTS model, we use two datasets respectively, a plain text dataset for CADEC style encoder and an audiobook dataset with (text, audio) pairs for Transformer TTS model.

Table 2: *classification results of five emotion categories on text-evaluation set*

Models	Overall accuracy (%)
BERT-style	57.63
CADEC-style	40.20
CADEC-style w/o. Stage 1	31.15
CADEC-style w/o. context	32.82

- **Text-training set and text-evaluation set.** For CADEC encoder, we collect a plain text dataset from e-books. It contains 1.7M conversation utterances with contexts. We split the dataset into three sets, 1,696,000 conversation utterances as **text-training set**, 2000 utterances as **in-domain text-evaluation set** for validation and 2000 utterances as **out-of-domain text-evaluation set** for testing.<sup>2</sup>
- **TTS-training set and TTS-evaluation set.** For TTS model, we apply an internal audio corpus as **TTS-training set**, read by a female speaker and a male speaker in Chinese. It is a complete 14 hours of expressive storytelling book, in which a portion of about 4 hours is conversation audio. Following most previous TTS work, we set in-domain test set from same audiobook and out-of-domain test set respectively to evaluate our model’s performance in different domains. We random select 10,981 utterances as training set, 200 utterances as **in-domain TTS-evaluation set** and 500 utterances from other books as **out-of-domain TTS-evaluation set**. Specifically, to evaluate voice quality of different genders, **out-of-domain TTS-evaluation set** contains both female and male speakers.

We use a supervised system that performed well in industrial as baseline. It contains two components: style prediction model and multi-style TTS model. The style prediction model (denoted as BERT-style) fine-tunes a pretrained BERT<sup>3</sup> with a downstream emotion classification task by leveraging annotated plain text dataset in a supervised manner. The annotated style tag is also applied when training baseline’s multi-style based Transformer TTS model.

In our proposed self-supervised style learning TTS framework, we use the same BERT backbone with content length of 256 as input in the same way as baseline BERT-style. In Stage 1 and Stage 2, we adopt Adam optimizer with mini-batch size of 32 and the initial learning rate as 1e-6. We firstly train 1,000 epoches with the contrastive loss to fine-tune BERT and autoencoder’s encoder in Stage 1 and continue to optimize all modules by minimizing  $\mathcal{L}_{total}$  in Stage 2 until convergence. By Grid Search, we set convergence threshold as 0.1%, and both the coefficient  $\beta$  and  $\gamma$  of the clustering loss as 0.5. For multi-style TTS system training, narrative utterance’s style embedding is zero-initialized in our proposed method. Both the baseline and our proposed method use similar settings including the acoustic feature as 80-dimensional log mel-spectrogram, window shift 12.5 ms. A MelGAN vocoder [28] of 24kHz is used for both.

#### 4.2. Subjective Evaluation of TTS

To evaluate the effectiveness of style embedding in TTS, we conduct subjective listening tests on Microsoft UHRS crowdsourcing platform. Participants are required to focus on speech

<sup>2</sup>It should be mentioned that for supervised baseline approach training and accuracy evaluation, all conversations utterances are annotated into five emotion categories (calm, joy, serious, fear and depressed) by crowdsourcing, but we do not use these annotations during the CADEC style encoder training.

<sup>3</sup><https://huggingface.co/bert-base-chinese>

style and expressiveness in all tests. For sentence-level speech, both MOS (Mean of Opinion Score) and CMOS (Comparative MOS) are conducted on **in-domain TTS-evaluation set** and **out-of-domain TTS-evaluation set**. Each audio is judged by at least 10 participants. As shown in Table 1, compared with the baseline, our model achieves better voice quality in terms of both MOS and CMOS, and importantly without any explicit annotation labels. The improvement in different datasets also demonstrates the effectiveness and robustness of our proposed method to cross-domain datasets of different genders.

In the audiobook scenario, a paragraph CMOS is used to evaluate the style expressiveness of multiple continuous utterances, as a paragraph or a session, in a conversation with narrative context. The judges are asked to rate audio considering context, which reflects the coherence and appropriateness of the audios’ expressiveness. We randomly select 20 paragraphs from **out-of-domain TTS-evaluation set** for each speaker with synthesized speech exceeding 30 seconds. The result shows significant preference over baseline, i.e. 0.22 for female and 0.23 for male speakers respectively. It verifies our assumption that style representation would be a continuous embedding other than a simple tag especially in a long context. Our model can express much more appropriate and diverse styles according to context, and achieves natural emotion transition between sentences.

#### 4.3. Analysis of Style Embedding

Compared with BERT-style that requires predefined categories, CADEC style embedding achieves relatively lower accuracy in emotion classification tasks (Table 2). In further analyzing text and human-annotated emotion labels, we find that five discrete labels can not model complex context-aware emotion representation, which leads to low classifier accuracy. However, as our proposed model performs prominently better than the baseline in TTS evaluation experiments, it demonstrates that continuous embedding is more suitable for TTS tasks. Meanwhile, removing Stage 1 leads to a significant drop in accuracy (Table 2) and implicit embedding space overlap (Figure 2(c)). This ablation study shows the importance of Stage 1 and further demonstrates that our proposed CADEC style embedding is effective in learning styles in addition to content. Moreover, removing text’s context reduces the overall accuracy by 7.18% which demonstrates the importance of context in style modeling. More experiment results and audio samples could refer to <https://wyh2000.github.io/InterSpeech2022/>

### 5. Conclusion and Future Work

In this work, we present a novel framework for self-supervised context-aware style encoding from unlabeled text for text-to-speech. Our proposed model has several advantages. 1) As a pre-trained style representation model from text, it can learn from massive unannotated data without requiring the corresponding audio. 2) by combining context-aware information and modeling style information in a continuous feature space, it can achieve natural expressiveness and emotion transition in long paragraph. In the future, we will enhance text style representation with a larger amount of text corpus for better accuracy and robustness. Besides, we will further explore mapping style embedding from text into acoustic feature space more robustly and guiding synthesized speech style in a flexible and controllable manner.

## 6. References

- [1] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 4693–4702. [Online]. Available: <https://proceedings.mlr.press/v80/skerry-ryan18a.html>
- [2] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *INTERSPEECH*. ISCA, 2018, pp. 3067–3071.
- [3] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 623–627.
- [4] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 5180–5189. [Online]. Available: <https://proceedings.mlr.press/v80/wang18h.html>
- [5] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 595–602.
- [6] T.-Y. Hu, A. Shrivastava, O. Tuzel, and C. Dhir, "Unsupervised style and content separation by minimizing mutual information for speech synthesis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 3267–3271.
- [7] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, A. Rosenberg, B. Ramabhadran, and Y. Wu, "Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 6699–6703.
- [8] S. Ma, D. McDuff, and Y. Song, "Neural tts stylization with adversarial and collaborative games," in *International Conference on Learning Representations*, 2018.
- [9] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer," *arXiv preprint arXiv:1711.05447*, 2017.
- [10] M. Kim, S. J. Cheon, B. J. Choi, J. J. Kim, and N. S. Kim, "Expressive Text-to-Speech Using Style Tag," in *Proc. Interspeech 2021*, 2021, pp. 4663–4667.
- [11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607. [Online]. Available: <https://proceedings.mlr.press/v119/chen20j.html>
- [12] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758.
- [13] Z. Yang, Y. Cheng, Y. Liu, and M. Sun, "Reducing word omission errors in neural machine translation: A contrastive learning approach," in *ACL*, 2019.
- [14] A. J. Bose, H. Ling, and Y. Cao, "Adversarial contrastive estimation," *CoRR*, vol. abs/1805.03642, 2018. [Online]. Available: <http://arxiv.org/abs/1805.03642>
- [15] E. Kharitonov, M. Rivière, G. Synnaeve, L. Wolf, P.-E. Mazaré, M. Douze, and E. Dupoux, "Data augmenting contrastive learning of speech representations in the time domain," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 215–222.
- [16] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, "Cdpam: Contrastive learning for perceptual audio similarity," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 196–200.
- [17] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 478–487. [Online]. Available: <https://proceedings.mlr.press/v48/xieb16.html>
- [18] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in *Ijcai*, 2017, pp. 1753–1759.
- [19] K. Do, T. Tran, and S. Venkatesh, "Clustering by maximizing mutual information across views," in *ICCV*. IEEE, 2021, pp. 9908–9918.
- [20] S. Park, S. Han, S. Kim, D. Kim, S. Park, S. Hong, and M. Cha, "Improving unsupervised image clustering with robust learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 278–12 287.
- [21] X. Zhan, J. Xie, Z. Liu, Y.-S. Ong, and C. C. Loy, "Online deep clustering for unsupervised representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6688–6697.
- [22] D. Zhang, F. Nan, X. Wei, S. Li, H. Zhu, K. R. McKeown, R. Nallapati, A. O. Arnold, and B. Xiang, "Supporting clustering with contrastive learning," in *NAACL-HLT*. Association for Computational Linguistics, 2021, pp. 5419–5430.
- [23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.
- [24] S. Buechel, S. Rücker, and U. Hahn, "Learning and evaluating emotion lexicons for 91 languages," in *ACL*. Association for Computational Linguistics, 2020, pp. 1202–1217.
- [25] J. X. Morris, E. Lifland, J. Y. Yoo, and Y. Qi, "Textattack: A framework for adversarial attacks in natural language processing," 2020.
- [26] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [27] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.
- [28] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," 2019.