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

Yihan Wu\textsuperscript{1,†}, Xi Wang\textsuperscript{2}, Shaofei Zhang\textsuperscript{2}, Lei He\textsuperscript{2}, Ruihua Song\textsuperscript{1,*}, Jian-Yun Nie\textsuperscript{3}

\textsuperscript{1} Gaoling School of Artificial Intelligence, Renmin University of China, China
\textsuperscript{2} Microsoft Azure Speech, China
\textsuperscript{3} Université de Montréal, China

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

\begin{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. \textsuperscript{1}

\textbf{Index Terms:} expressive TTS, self-supervised learning, deep clustering, representation learning, contrastive learning
\end{abstract}

\section{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.

\section{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 \([13, 14, 15, 16]\).
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 \( D = \{U_i\}_{i=1}^D \) where \( U_i = \{C_i^{-}, u_i, C_i^{+}\} \) represents text utterance \( u_i \) and its context. \( C_i^{-} = \{u_{i-1}, \ldots, u_{i-k}\} \) is the preceding utterances of \( u_i \) and \( C_i^{+} = \{u_{i+k+1}, \ldots, u_{i+k}\} \) is the following utterances of \( u_i \). Our goal is to learn a style encoding model \( s_i = g(U_i) \) from \( D \) 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 (valence, 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:

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

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:

\[
    h_i = MLP(r_i)
\]

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

<table>
<thead>
<tr>
<th>Metrics</th>
<th>MOS</th>
<th>CMOS</th>
<th>Paragraph CMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Settings</strong></td>
<td><strong>Recording</strong></td>
<td><strong>Baseline</strong></td>
<td><strong>Our Model</strong></td>
</tr>
<tr>
<td>in-domain</td>
<td>4.35 ± 0.01</td>
<td>4.26 ± 0.07</td>
<td>4.34 ± 0.06</td>
</tr>
<tr>
<td>out-of-domain</td>
<td>Female</td>
<td>4.37 ± 0.1</td>
<td>4.21 ± 0.07</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>4.35±0.13</td>
<td>4.1±0.1</td>
</tr>
</tbody>
</table>

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 $D$, we randomly sample a minibatch data $B = \{ U_i \}_{i=1}^N$, and generate its augmented data $\tilde{B} = \{ \tilde{U}_i \}_{i=1}^N$, where $U_i = \{ C_{R_i}^+, \tilde{u}_i, C_{R_i}^- \}$, $U_i$ and $\tilde{U}_i$ are treated as positive pairs while the other $N-1$ pairs $(< U_i, \tilde{U}_k >)_{i \neq k}$ 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

$$L_{\text{contrastive}} = \frac{1}{N} \sum_{i=1}^{N} L_i$$

Here $\tau$ is the temperature parameter and $\delta_{x \neq y}$ is the indicator function. The contrastive loss for a minibatch is computed by averaging over all instances in $B$ and its augmented data $\tilde{B}$:

$${L}_{\text{contrastive}} = \frac{1}{N} \sum_{i=1}^{N} L_i$$

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

$$L_{\text{clustering}} = KLD(P||Q) = \sum_{k} \sum_{i} p_{ik} \log \frac{p_{ik}}{q_{ik}}$$

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 $q_{ik}$.

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

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

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

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

$$L_{\text{reconstruction}} = \sum_{i=1}^{N} \| p_i - r_i' \|^2_2$$

Therefore, the objective of Stage 2 is defined as

$$L_{\text{total}} = L_{\text{contrastive}} + \beta L_{\text{clustering}} + \gamma L_{\text{reconstruction}}$$

We tune all trainable parameters to optimize $L_{\text{total}}$ 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

<table>
<thead>
<tr>
<th>Models</th>
<th>Overall accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-style</td>
<td>37.63</td>
</tr>
<tr>
<td>CADEC-style</td>
<td>40.20</td>
</tr>
<tr>
<td>CADEC-style w/o. Stage 1</td>
<td>31.15</td>
</tr>
<tr>
<td>CADEC-style w/o. context</td>
<td>32.82</td>
</tr>
</tbody>
</table>

- **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. 2

- **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 randomly 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.

In this work, we present a novel framework for 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 epochs with the contrastive loss to fine-tune BERT and autoencoder, 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. 2

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 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.

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’ 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://wzh2000.github.io/InterSpeech2022/
6. References


