SYNCLR: A SYNTHESIS FRAMEWORK FOR CONTRASTIVE LEARNING OF OUT-OF-DOMAIN SPEECH REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning generalizable speech representations for unseen samples in different domains has been a challenge with ever increasing importance to date. Although contrastive learning has been a prominent class of representation learning approaches, the state-of-the-art (SOTA) contrastive learning methods were found to have limited ability for learning unseen out-of-domain speech representations. This paper presents SynCLR, a synthesis framework for contrastive learning of speech representations that can be generalized over unseen domains. Specifically, instead of using data augmentation approach, SynCLR employs data synthesis for multi-view generation. To ensure a highly-varied conditional speech distribution in view generation, we design a novel diffusion-based speech synthesizer. A new contrastive loss is also proposed to construct multiple embedding spaces, each of which preserves view-sensitive information to reduce domain reliance for a better disentanglement. Our experiments showed that SynCLR outperformed the SOTA contrastive learning methods with a 21.1% relative reduction of EER in speaker verification tested on an unseen speech corpus, and considerably reduced 50.8% relative FIDs in a challenging speech-to-image translation task given out-of-domain test speeches.

1 INTRODUCTION

Learning representations of out-of-domain speeches, i.e., speeches from speaker, speaking style, or content unseen in training, is a challenging problem that heavily relies on the model's generalization ability. Towards learning generalized representations, a majority of the prior works (Oord et al., 2018; Tian et al., 2020) studied contrastive methods, which are designated to pull the positive sample pairs together and push the negative ones apart in the high-dimensional embedding space. The recent state-of-the-art contrastive learning methods (Chen et al., 2020b; Al-Tahan & Mohsenzadeh, 2021) shared a common idea that employs data augmentation for view generation (Bachman et al., 2019; Tian et al., 2020). It has been shown that view generation is useful for enhancing the robustness of the learned representation since diverse views can lead to a better exploration of the data distribution.

However, the aforementioned contrastive learning methods are limited in their generalizability of representations for out-of-domain speeches mainly because of two challenges: (1) The augmented data distribution has limited diversity. Apparent degradation still emerges when encountering speech with content, prosody or speaker unseen in training. (2) Projecting data from different views into one single space has been demonstrated to obtain representations invariant to transformations, yet this also makes the learning of view-sensitive and disentangled representations difficult.

To address the above challenges, we introduce a novel synthesis framework called SynCLR for contrastive learning of out-of-domain speech representations. The unique advantages of SynCLR can be summarized into three points: (1) Multi-view data synthesis was used in place of data augmentation for task-dependent view generation, which provides diverse views to conduct generalizable representation learning. (2) To avoid mode collapse (Creswell et al., 2018) in the dominated GAN-based generative models, which leads to very similar output samples from a single or few modes of the distribution, especially in the strongly conditional generation task (e.g., speech synthesis), we designed a diffusion-based synthesizer named SynGrad. SynGrad can be used to efficiently generate

view-conditional speech samples, and has shown a greater synthesis variability, which is essential for improving the robustness of speech representations. (3) To reduce the limitation of speech encoder applied to a specific domain, SynCLR jointly optimizes multiple view-sensitive contrastive objectives to relate the projected embedding spaces to the input views.

Our proposed SynCLR approach was evaluated on a self-supervised speaker verification task and a challenging supervised speech-to-image generation task to assess the efficiency and the generalization ability of SynCLR in learning out-of-domain speech representations. Finally, SynCLR outperformed the SOTA contrastive learning methods with a 17.2% relative improvement of EER in verifying unseen speakers, and considerably reduced 50.8% relative FID in a challenging speech-to-image translation task given out-of-domain test speeches. The experimental results demonstrated that our proposed SynCLR framework can significantly improve the out-of-domain speech representation learning.

Our main contributions are summarized below:

- We introduce a multi-view synthesizer in the contrastive learning framework to generate diverse yet view-controllable samples, which can improve the generalization power.
- We design a fast and high-quality waveform generative model SynGrad. SynGrad avoids mode collapse in previous GAN-based methods and succeeds to expand data distribution for multi-view data synthesis.
- We propose a novel contrastive learning objective to construct multiple embedding spaces, each of which preserves view-sensitive information to reduce domain reliance for a better disentanglement.

2 Background: A Simple Contrastive Learning Framework

Contrastive learning is one of the prominent self-supervised learning approaches that learn a representation by grouping similar data pairs and repelling dissimilar pairs on a high-dimensional embedding space. It has been shown successful for learning rich representations in the speech domain. Recently, Chen et al. (2020b) introduced a simple contrastive learning framework called *SimCLR*, which produced remarkable results outperforming previous self-supervised and semi-supervised representation learning methods. The SimCLR framework comprises four major components:

- A stochastic data augmentation operator $\mathcal{D}$ that generates augmented samples from a reference data sample along different views, e.g., varying speakers, prosodies, or transcripts.
- An encoder $f : \mathcal{X} \mapsto \mathcal{H}$ that extracts a hidden representation $\boldsymbol{c} \in \mathcal{H} \subseteq \mathbb{R}^D$ from a given speech waveform $\boldsymbol{x} \in \mathcal{X}$.
- A projector $h : \mathcal{H} \mapsto \mathcal{Z}$ that further maps an extracted speech feature into an embedding space $\mathcal{Z} \subseteq \mathbb{R}^N$ for calculating the contrastive loss.
- A contrastive loss ℓ that evaluates the similarity of a positive pair of data examples relative to n negative pairs on the embedding space.

We specifically define the contrastive learning framework for the speech domain. Let $\boldsymbol{x}_0$ be the reference speech, we denote an augmented speech as $\tilde{\boldsymbol{x}}_0^{(v)} \sim \mathcal{D}^{(v)}(\boldsymbol{x}_0)$, where v indexes an augmentation view. Then, considering other speech references $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, we can form a positive data pair $(\boldsymbol{x}_0, \tilde{\boldsymbol{x}}_0^{(v)})$ together with n negative pairs $(\boldsymbol{x}_1, \tilde{\boldsymbol{x}}_1^{(v)}), ..., (\boldsymbol{x}_n, \tilde{\boldsymbol{x}}_n^{(v)})$ along the v-th augmentation view. After the positive and negative pairs are obtained, the contrastive loss can be computed. A common choice of ℓ is the InfoNCE loss (Oord et al., 2018):

$$\ell_{\text{infoNCE}}(v) := -\log \frac{\exp\left(\boldsymbol{z}_0{}^\top \boldsymbol{W}_v \tilde{\boldsymbol{c}}_0^{(v)}\right)}{\exp\left(\boldsymbol{z}_0{}^\top \boldsymbol{W}_v \tilde{\boldsymbol{c}}_0^{(v)}\right) + \sum_{i=1}^{n} \exp\left(\boldsymbol{z}_0{}^\top \boldsymbol{W}_v \tilde{\boldsymbol{c}}_i^{(v)}\right)}, \tag{1}$$

where $\boldsymbol{z}_i = f(\boldsymbol{x}_i)$, $\boldsymbol{W}_v$ is a learnable matrix defined for the view v. Here and below, we define the shorthand, $\tilde{\boldsymbol{z}}_i^{(v)} = f(\tilde{\boldsymbol{x}}_i^{(v)})$ and $\tilde{\boldsymbol{c}}_i^{(v)} = h(\tilde{\boldsymbol{z}}_i^{(v)})$. Note that Oord et al. (2018) considered different v as different step numbers ahead of the current time index, but we can generally consider them as different views as in (Bachman et al., 2019; Tian et al., 2020).

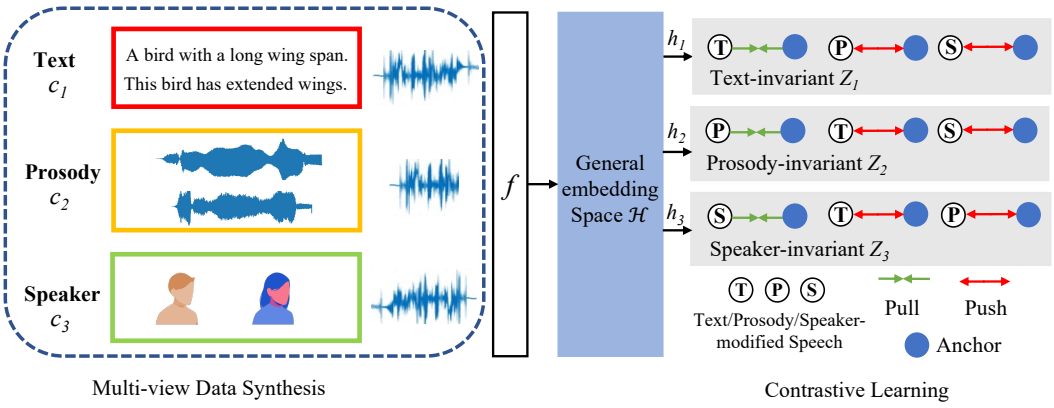


Figure 1: A block diagram of the SynCLR framework. SynCLR learns multiple embedding spaces to construct view-sensitive information. Take text-invariant embedding z_1 as an example: The text-modified synthetic sample is considered to be positive while the rest of the samples in the batch are considered negative samples, and thus z_1 represents the text-invariant information generalizable to text-sensitive out-of-domain distribution.

Alternatively, Chen et al. (2020b) employed a normalized temperature-scaled cross entropy loss (NT-Xent), which can be computed given a minibatch of n reference audios. Let $s(\boldsymbol{u}, \boldsymbol{v}) := \boldsymbol{u}^\top \boldsymbol{v} / \max(\|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2, \epsilon)$ with a small positive constant ϵ. Considering two different augmentation views v and v', the n references are transformed into $2n$ augmented samples. Then, we compute

$$\ell_{\text{NT-Xent}}(i, v, v') := -\log \frac{\exp\left(s(\tilde{\boldsymbol{z}}_i^{(v)}, \tilde{\boldsymbol{z}}_i^{(v')})/\tau\right)}{\exp\left(s(\tilde{\boldsymbol{z}}_i^{(v)}, \tilde{\boldsymbol{z}}_i^{(v')})/\tau\right) + \sum_{j \neq i} \exp\left(s(\tilde{\boldsymbol{z}}_i^{(v)}, \tilde{\boldsymbol{z}}_j^{(v')})/\tau\right)}, \tag{2}$$

where τ is a hyperparameter controlling the temperature. In essence, SimCLR aims at learning efficient visual representations by maximizing agreement between differently augmented views of the same data and maximizing difference across contrasting images. From a different aspect, Xiao et al. (2020) proposed to capture varying and invariant factors for constructing separate embedding spaces.

3 SynCLR: A Synthesis Contrastive Learning Framework

In this paper, we propose a novel synthesis framework for contrastive learning of multi-domain speech representations called *SynCLR*.

Compared with the SOTA contrastive framework described before, SynCLR has the following key differences in learning speech representations: (1) To overcome the limitation of diversity in data augmentation, we propose a multi-view data synthesis strategy for speech representation learning unseen domain. SynCLR does not require a reference sample to generate augmented samples for contrastive learning, on the contrary, the synthesizer pre-defines the conditions to synthesize a pair of positive or negative samples for computing the contrastive loss. (2) For high-quality and diverse speech synthesis, we introduce a novel diffusion probabilistic model named SynGrad, which guarantees stable training and avoids mode collapse in dominant GAN-based methods. (3) Towards learning generalizable, view-sensitive and disentangled representation, we introduce a multi-head embedding learning with a novel contrastive loss.

Overall, the SynCLR training procedure can be mainly divided into three stages: (i) Multi-view data synthesis; (ii) Multi-head embedding learning; and (iii) Contrastive loss calculation. Each of these stages is described in the following sections. We illustrate the SynCLR framework in Figure 1.

3.1 MULTI-VIEW DATA SYNTHESIS

As data diversity is vital for generalizable representation learning, SynCLR employs a multi-view data synthesis approach for data manipulation, denoted as a $\mathrm{mani}(\cdot)$ function. By modifying the data manipulation strategy in the contrastive learning framework, we are able to explore a more diverse data distribution, and hence improving the generalizability of the learned representations.

Generative Model In the speech domain, high-quality, expressive and customized text-to-speech syntheses (TTS) have been proven successful, which provides an ideal environment for flexible multi-view contrastive learning. Existing methods are dominated by the generative adversarial networks (GANs), which, however, has been criticized for mode collapse and limited sample diversity (Dhariwal & Nichol, 2021; Creswell et al., 2018). To encounter stable training and broaden data distribution of synthetic samples, we employ the state-of-the-art generative model – denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020), which have been shown to surpass GANs in terms of sampling quality (Dhariwal & Nichol, 2021). Yet, in order to maintain high-quality synthesis, DDPMs require hundreds to thousands of steps to generate a sample (Chen et al., 2020a; Kong et al., 2020b). To alleviate the efficiency issue, a recent work proposed bilateral denoising diffusion models (BDDMs) (Lam et al., 2021a) to speed up DDPMs into more than 10x meanwhile remaining high-fidelity speech synthesis. BDDMs mainly consist of a score network θ and a scheduling network ϕ. The detailed formulation of BDDMs is presented in Appendix A.

Formally, given a generative diffusion model parameterized by a score network θ, we synthesize a speech sample given a set of V conditions $\mathbb{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_V\}$ corresponding to V views, e.g., text, prosody, and speaker: $\boldsymbol{x} \sim \mathrm{Diff}_\theta(\boldsymbol{\beta}, \mathbb{C})$, where Diff_θ is the pre-trained BDDM for inference, and $\boldsymbol{\beta} \in \mathbb{R}^S$ is the noise schedule for sampling. Figure 1 shows an example of having three views, and $\{\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3\}$ denote the values in the text view, the prosody view, and the speaker view, respectively.

Model Architecture The architecture of the proposed end-to-end diffusion synthesizer is shown in Figure 2, which consists of a text-to-spectrogram model FastSpeech 2 (Ren et al., 2020) and an efficient waveform generator named SynGrad. Inspired by the recent works (Ren et al., 2020; Kim et al., 2020) in controllable speech synthesis, we convert the input text into a sequence of hidden variables and add variance information (i.e., duration, pitch, and energy) to generate mel-spectrograms with multiple speech views. Follow from this, SynGrad converts the mel-spectrograms into speeches with a noise scheduler (Lam et al., 2021a)

To provide diverse views in our contrastive learning framework, in FastSpeech 2, we employ duration and pitch predictors, and embed the speaker identities into the encoded text sequence for generating the multi-speaker view of speech.

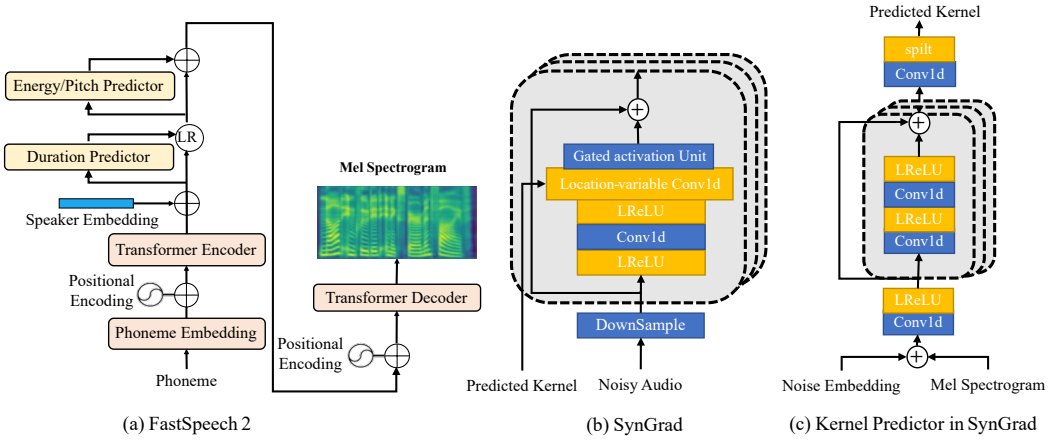
In SynGrad, we adopt kernel predictors for location-variable convolution (Jang et al., 2021; Zeng et al., 2021), which can capture the conditional local information, leading to a much more efficient sampling process than those with other existing architectures, e.g., WaveGrad (Chen et al., 2020a) and DiffWave (Kong et al., 2020b). Specifically, the kernel predictors directly learn multiple sets of convolution kernels according to the diffusion step embedding and the synthesized spectrogram features. These kernels can be used to perform convolution operations on the associated intervals in the input sequence, which is superior to the traditional convolution networks in modeling the long-term waveform dependencies.

Noise Scheduling Acceleration The sampling quality and speed of a diffusion probabilistic model are directly related to the pre-defined noise schedule for sampling $\boldsymbol{\beta}$. Ho et al. (2020) used a linear noise schedule for sampling but it is prohibitive for efficient sampling. To solve this problem, we follow Lam et al. (2021a) to use the GALR (Lam et al., 2021b) network GALR_ϕ for noise scheduling, i.e., starting from the hyperparameters (α_S, β_S), we recursively estimate β_s as follows:

$$\beta_s = \mathrm{GALR}_\phi(\boldsymbol{x}_s, \alpha_{s+1}, \beta_{s+1}), \tag{3}$$

$$\boldsymbol{x}_{s-1} \sim \mathcal{N}\left(\frac{1}{\sqrt{1-\beta_s}}\left(\boldsymbol{x}_s - \frac{\beta_s}{\sqrt{1-\alpha_s^2}}\epsilon_\theta\left(\boldsymbol{x}_s, \alpha_s\right)\right), \frac{1-\alpha_{s-1}^2}{1-\alpha_s^2}\beta_s \boldsymbol{I}\right), \tag{4}$$

where $\alpha_s = \alpha_{s+1}/\sqrt{\beta_{s+1}}$, $\alpha_{s-1} = \alpha_s/\sqrt{\beta_s}$ and ϵ_θ is a pre-trained score network. The predicted noise schedule is then mapped to discrete step indices following (Kong et al., 2020b) for the score



(a) FastSpeech 2 (b) SynGrad (c) Kernel Predictor in SynGrad

Figure 2: The pipeline of multi-view data synthesis. $\oplus$ denotes the element-wise addition operation, and LR denotes the length regulator.

network conditioning on a discrete step index. In the following context of SynCLR, all $\boldsymbol{\beta}$s are assumed produced as the above.

3.2 Multi-head Embedding Learning

As described above, data synthesis contributes to expanding the data distribution and improving the generalizability of learned speech representation. To extract the representation vectors from speeches, we used a neural network based encoder $f : \mathcal{X} \mapsto \mathcal{H}$, from which we obtain the general feature $\boldsymbol{v}_i$ of a synthesized speech $\boldsymbol{x}_i$ by mapping it into a D-dimensional embedding space $\mathbb{R}^D$. Then, a v-dependent projector $h_v : \mathcal{H} \mapsto \mathcal{Z}$ further maps the extracted general representations into an embedding space $\mathcal{Z}_v$ corresponding to the v-th view conditioning the speech synthesis. Here, the number of dimensions in different view-corresponded embedding spaces can also be different: $\mathcal{Z}_v \subseteq \mathbb{R}^{D_v}$ for $v = 1, \ldots, V$. Notably, the definition of projector in SynCLR is different from previous works, which projected every view into a single embedding space that is invariant to all augmentations (Oord et al., 2018; Chen et al., 2020b), or projected every view into different embedding sub-spaces that hold no direct correspondence to the V views (Xiao et al., 2020).

3.3 Contrastive Loss

Last but not least, we compute the contrastive loss. Other than defining positive or negative samples based on augmentation, the SynCLR framework synthesizes contrastive samples by defining varying and invariant views for the generative model as different conditions.

Formally, we define a batch of n collections of conditions $\{\mathbb{C}_i\}_{i=1}^n = \{\{\boldsymbol{c}_{i,1}, \ldots, \boldsymbol{c}_{i,V}\}\}_{i=1}^n$ for synthesis, where $\boldsymbol{c}_{i,v}$ denotes the condition values for the v-th view in the i-th collection of the batch. Based on these conditions, we synthesize the reference samples $\boldsymbol{x}_i \sim \text{Diff}_\theta(\boldsymbol{\beta}, \mathbb{C}_i)$ for $i = 1, \ldots, n$. For defining the contrastive samples relative to the reference samples, we vary a specific view v to obtain $\mathbb{C}_{i\setminus v} = \{\boldsymbol{c}_{i,1}, \ldots, \boldsymbol{c}_{j,v}, \ldots, \boldsymbol{c}_{i,V}\}$, which implies replacing the v-th element $\boldsymbol{c}_{i,v}$ of $\mathbb{C}_i$ by a different condition $\boldsymbol{c}_{j,v}$ from a randomly selected collection $\mathbb{C}_j$ in the same batch with $j \neq i$. Followed from this, we define a different set of n synthesized samples each with the v-th view varied: $\boldsymbol{x}_i^{(v)} \sim \text{Diff}_\theta(\boldsymbol{\beta}, \mathbb{C}_{i\setminus v})$ for $i = 1, \ldots, n$.

Given the synthesized speech samples, for each v, we compute n pairs of embedding vectors: $\{\boldsymbol{z}_{i,v}, \boldsymbol{z}_i^{(v)}\}_{i=1}^n$, where $\boldsymbol{z}_{i,v} = h_v(f(\boldsymbol{x}_i))$ are computed from the reference samples, and $\boldsymbol{z}_i^{(v)} =$

$h_v(f(\boldsymbol{x}_i^{(v)}))$ are computed from the view-varied samples. Then, we define the contrastive loss as

$$\ell_{\text{SynCLR}}(i) := -\sum_{v=1}^{V} \log \frac{\exp\left(s(\boldsymbol{z}_{i,v}, \boldsymbol{z}_i^{(v)})/\tau\right)}{\exp\left(s(\boldsymbol{z}_{i,v}, \boldsymbol{z}_i^{(v)})/\tau\right) + \sum_{j \neq i} \exp\left(s(\boldsymbol{z}_{i,v}, \boldsymbol{z}_j^{(v)})/\tau\right)}, \tag{5}$$

where, similar to Eq. 2, $s(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \max(\|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2, \epsilon)$ with a small positive constant ϵ, and τ is a hyperparameter controlling the temperature.

Both the encoder network $f(\cdot)$ and the projectors $h_1(\cdot), \ldots, h_V(\cdot)$ are trained with the above defined contrastive loss $\ell_{\text{SynCLR}}(i)$ by sampling i uniformly at each training step. After the training is completed, given an unseen speech sample $\boldsymbol{x}^\star$, we use the concatenation of V-views embeddings: $\text{Concat}([h_1(f(\boldsymbol{x}^\star)), ..., h_V(f(\boldsymbol{x}^\star))])$ as the speech representation.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

A majority of representation learning methods adopt a pretraining stage for expanding the general data distribution and then finetuning to the downstream datasets. Different from these works, SynCLR does not follow a pretraining & finetuning pipeline, instead, it can be directly trained on a specific task to learn speech representations generalizable to unseen test cases without using the labels for that task. The main idea is to correspondingly define the view spaces for a given task to bring task-specific improvements. Similar to existing works (Ye et al., 2021; Nasiri & Hu, 2021; Seshadri & Lerch, 2021), the contrastive learning approach could be incorporated into the existing downstream model to improve their performance. Similarly, we evaluated SynCLR on two downstream tasks (i.e., speaker verification, speech-to-image translation) to test our hypotheses.

Multi-view data synthesis Our multi-view data synthesis pipeline consists of several neural networks for promoting quality, including a "variance adaptor" FastSpeech 2, a diffusion probabilistic waveform generative model SynGrad, and a scheduling network GALR for scheduling acceleration.

We employed both single-speaker LJSpeech (Ito & Johnson, 2017) and multi-speaker LibriTTS (Zen et al., 2019) datasets to train the speech synthesizer for view generation. For a given text input, we generate corresponding speeches by conditioning the speech synthesizer on multiple views, e.g., prosody and speaker. In judgment for generalization of speech representation learning model, we prepare both in-domain and out-of-domain testing sets. More details have been attached in Appendix E.

Learnt representations Taken time-frequency audio features (i.e. spectrograms) as input, the speech representation learning network f_{enc} consists of a two-layer 1-D convolution block and a LSTM layer. It generates 1024 dimensional general embeddings from log Mel filter bank spectrograms. In this work, we use 2-layer MLP projection heads f_{proj} to project the general representation to content, prosody and speaker embedding spaces.

4.2 SPEAKER VERIFICATION

Motivation To investigate the effect of SynCLR acting as a contrastive learning approach which could be incorporated into the existing downstream models, we conducted experiments on speaker verification.Speaker verification (Wan et al., 2018; Lai, 2019) aims to verify whether the speakers of a pair of utterances match. The pair may not be presented beforehand, and hence it is a challenging open-set problem. Furthermore, testing on more **noisy** and **diverse** multi-speaker out-of-domain speech samples is a promising path to evaluate model generalization, so that we prepared unseen test samples from VoxCeleb1 to constitute the out-of-domain testing set.

Implementation details We implemented competitive contrastive learning approaches for comparison and trained speech encoder on the train-clean-100 subset of LibriTTS. Detailed configurations are presented in Appendix F. Cosine similarity served as a back-end scoring method during testing. The evaluation results are assessed in equal error rates (EER).

Quantitative Results Experimental results are shown in Table 1. Compared with the competing contrastive learning approaches, our proposed SynCLR mechanism captured more discriminative representations and reduced EER by 5.7%, 4.34%, and 3.01% relative to the downstream model (i.e., d-vectors), the model incorporated with InfoNCE, and SimCLR contrastive learning approaches during out-of-domain testing, respectively. The proposed SynCLR had been demonstrated to be superior to previous contrastive learning objectives and preserve better generalization in out-of-domain distribution.

Method	In-domain	EER (↓)	Out-of-domain	EER (↓)
d-vectors	LibriTTS	8.23	VoxCeleb1	32.05
+infoNCE	LibriTTS	8.16	VoxCeleb1	30.69
+SimCLR	LibriTTS	8.02	VoxCeleb1	29.36
+SynCLR	LibriTTS	**7.55**	VoxCeleb1	**26.35**

Table 1: Quantitative results of SynCLR and competitive contrastive approaches on speaker verification.

4.3 SPEECH-TO-IMAGE TRANSLATION

Motivation Speech-to-image synthesis (Li et al., 2020; Wang et al., 2021) generates images that have semantic contents corresponding to the input speech descriptions. The performance of conventional speech-to-image translators, to a large extent, depends on speech representations, which challenges a successful disentanglement of view-sensitive information to reduce domain reliance. This thus provides a touchstone to evaluate and compare different speech representation learning methods. What's more, we consider an even more challenging but realistic scenario, where out-of-domain test sets to assess the generalization abilities of different methods. During the out-of-domain evaluation, distinctively trained speech encoders were fed with speeches with text, prosody, or speaker unseen in training.

Implementation details Several contrastive learning methods were taken involved in speech representation learning. Given speech representation, we chose two popular GAN-based speech-to-image synthesis models (Li et al., 2020; Wang et al., 2021) to generate an image in the second stage. During testing, we adopted quantitative evaluation metrics of Inception Score (IS) (Heusel et al., 2017) and Fréchet Inception Distance (FID) (Salimans et al., 2016) both in-domain and out-of-domain testing sets. Detailed configurations and definitions are presented in Appendix G.

Quantitative Results The experimental results are shown in Table 4.3. Rich and high-fidelity out-of-domain speech guidance is learned through proposed synthesis contrastive learning framework. Compared with the two baselines, SynCLR improved the quality of synthetic images in terms of both IS and FID. In out-of-domain speech-to-image translation, IS was improved from 3.73 to 5.35 with DirectGAN and from 4.80 to 5.33 with S2IGAN. For the FID metric, SynCLR also improved the baseline DirectGAN and S2IGAN significantly by 50% and 23% in out-of-domain image synthesis, respectively. As the data distribution changed in unseen test set, we can witness an increase of the inception scores compared to in-domain evaluation.

Method	In-domain		Out-of-domain	
	FID(↓)	IS(↑)	FID(↓)	IS(↑)
DirectGAN	15.48	4.96	29.76	3.73
+SimCLR	14.55	4.98	17.81	5.03
+SynCLR	**11.55**	**5.23**	**14.65**	**5.35**
S2IGAN	15.46	4.82	15.98	4.80
+SimCLR	14.77	4.76	15.43	5.17
+SynCLR	**11.55**	**4.97**	**12.26**	**5.33**

Table 2: Quantitative results of proposed SynCLR method on speech-to-image generation. We conduct experiments on two baselines DirectGAN and S2IGAN.

In summary, the quantitative experimental results demonstrated that our proposed framework SynCLR is efficient in learning out-of-domain speech representations and showed superiority in guiding conditional image generation.

Qualitative Findings To further compare our proposed approach with the baselines, we attached the synthetic images guidance in challenging out-of-domain speeches. Following the generated samples in figure 3, we have two observations: (1) For expressiveness, our method better highlighted the main subject of the image from its background. As speech representations became more abundant,

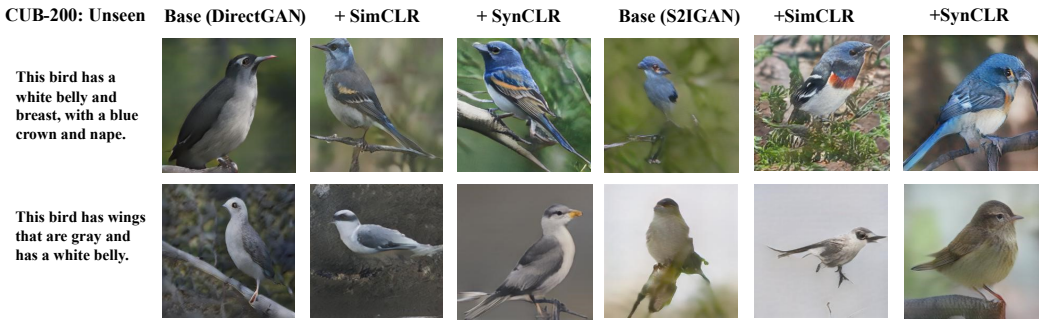


Figure 3: Comparison of synthesized images over out-of-domain CUB dataset.

more details in synthesized images emerged. Generated depictions seemed to be dull in baselines, while the ones guided by adding SynCLR's learned representations in column 1^{st} and 6^{th} were more colorful and vivid. (2) For realism, our approach better matched with the speech descriptions and kept correspondence in most cases. For instance, the synthesized image in column 1^{st} of first utterance did not represent corresponding text "blue crown", whereas the synthesized image generated by adding SynCLR's learned representations produced relevant images. Contrasting speech representations contributes a lot to better understand what people say even in out-of-domain scenarios. Additional image samples are presented in Appendix G.

4.4 Ablation Study

Multi-View Generation We conduct an ablation study of multiple views and embedding spaces in contrastive learning. To intuitively present respective influences on speech representation, we experiment on downstream speech-to-image generation with S2IGAN. The results are listed in Table 3, and we have two observations:

Method	Variance View			In-domain		Out-of-domain	
	T	P	S	FID(↓)	IS(↑)	FID(↓)	IS(↑)
SimCLR	✓			14.77	4.76	15.43	5.17
SimCLR	✓	✓		13.75	4.81	18.39	5.22
SimCLR	✓	✓	✓	11.69	4.88	14.96	5.21
SynCLR	✓	✓	✓	**11.55**	**4.97**	**12.26**	**5.33**

Table 3: Ablation study results of multiple views and embedding spaces in contrastive learning. T, P, and S denote text, prosody, and speaker.

(1) With the increasing number of variance views in SimCLR, the quality of synthesized images guided by both in-domain and out-of-domain speeches could be boosted. This finding proves that expanding data distribution through multi-view generation plays a vital role in multi-domain representation learning. Note that the additional view of prosody hurts performance in terms of FID for out-of-domain testing mainly beacuse of the potential collapse of generative adversarial networks. (2) Even encountered with similar data diversity, SynCLR achieves a 17.2% relative decreased of FID degradation in out-of-domain speeches compared with the SOTA contrastive framework. It has been demonstrated that novel contrastive loss objective with disentangled and view-sensitive embedding effectively contributes in learning out-of-domain speech representations.

Deep Generative Model For the reason that FastSpeech 2 plays a fundamental role in including variances for multiple views generation, we mainly conduct experiments on the neural vocoding stage with competing architecture including WaveGrad (Chen et al., 2020a), Diffwave (Kong et al., 2020b) and HiFi-GAN (Kong et al., 2020a). Vocoders are conditioned on the mel-spectrograms computed from ground truth audio during training. GitHub implementations are used for reproducibility and the configurations follow their original papers. RTF denotes the real-time factor, that the seconds required for the vocoder to synthesize one-second audio conditioned on mel-spectrogram synthesized by FastSpeech 2. We measure RTF on

Model	MOS (↑)	RTF (↓)
FastSpeech 2 + HiFi-GAN	3.91±0.31	**0.007**
FastSpeech 2 + Diffwave	3.66±0.06	0.219
FastSpeech 2 + WaveGrad	4.00±0.09	0.230
Ours (FastSpeech 2 + SynGrad)	**4.24±0.04**	0.030

Table 4: Evaluation on multi-view data synthesis with competing architecture.

an NVIDIA V100 GPU, and all diffusion-based neural vocoders generate samples within 6 reverse steps.

For easy comparison of audio quality and synthesis speed, the results are compiled and presented in Table 4, audio samples are available in: `https://synclr.github.io/`. As a summary, we have the following key findings: (1) For audio quality, SynGrad beat the strong GAN-based vocoder HIFI-GAN in terms of speech quality, which at the same time outperformed the previous diffusion-based generative model with short reverse steps (i.e., 6 steps) in terms of denoising ability. (2) For inference speed, SynGrad led to 7.2x and 7.6x speed up with respect to diffusion models including Diffwave and WaveGrad. In conclusion, the experiment result demonstrated the robustness and superiority of SynGrad for fast and high-fidelity multi-view data synthesis. More ablation study results are presented in Appendix E.3.

5 Related Works

5.1 Contrastive Learning

Contrastive learning approaches learn representations by contrasting positive pairs against the negative ones. Contrastive Predictive Coding (Oord et al., 2018) was introduced as a universal unsupervised learning approach to extract useful representations from high-dimensional data. With superior results, Chen et al. (2020b) presented a simple framework called SimCLR for an effective contrastive learning of visual representations. Khosla et al. (2020) extended the self-supervised batch contrastive approach to the fully-supervised setting. Al-Tahan & Mohsenzadeh (2021) then extended SimCLR to learn better speech representations. Distinctively, this paper presents a novel contrastive learning framework that improves generalization ability through the multi-view data synthesis and the multi-head embedding learning for speech representation learning.

5.2 Domain Generalization

The goal of domain generalization is to learn domain-invariant representations using only the training data from the source domains. Different from unsupervised domain adaptation, target domain data is inaccessible during the training, making the task more challenging. Our SynCLR framework can be viewed as a self-supervised solution to the domain generalization problem. Among recent successful methods for domain generalization, MuST (Ghiasi et al., 2021) adopt self-training to aggregate labeled and unlabeled training data to learn general feature representations. Lorincz et al. (2021) learned a better speaker identity representation by introducing an additional loss. Li et al. (2018) simulated the meta-train and meta-test tasks in training domains to enhance the performance. Instead of applying the aforementioned techniques, we include a contrastive learning framework in generalization to out-of-domain speeches.

6 Conclusion

In this paper, we presented the success of SynCLR in learning out-of-domain speech representations. To improve the model generalization ability over unseen speeches, SynCLR adopted a speech synthesizer to provide diverse views of data distribution. For efficiently generating view-conditional speech samples without mode collapse, we designed a diffusion-based synthesizer named SynGrad. In SynCLR, we define multi-head embedding spaces, in which the model learns view-disentangled representations to promote out-of-domain generalization. Finally, the consistent and significant empirical results suggested that SynCLR is superior to the previous SOTA contrastive learning methods when testing with out-of-domain samples in both the speaker verification and the challenging speech-to-image translation tasks. All in all, the proposed SynCLR framework leads to a promising path towards an advanced contrastive learning of out-of-domain speech representations.

REFERENCES

Haider Al-Tahan and Yalda Mohsenzadeh. Clar: Contrastive learning of auditory representations. In *International Conference on Artificial Intelligence and Statistics*, pp. 2530–2538. PMLR, 2021.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020b.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35 (1):53–65, 2018.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. Exploring wav2vec 2.0 on speaker verification and language identification. *arXiv preprint arXiv:2012.06185*, 2020.

Golnaz Ghiasi, Barret Zoph, Ekin D Cubuk, Quoc V Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. *arXiv preprint arXiv:2108.11353*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

Keith Ito and Linda Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. *arXiv preprint arXiv:2106.07889*, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *arXiv preprint arXiv:2005.11129*, 2020.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *arXiv preprint arXiv:2010.05646*, 2020a.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020b.

Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pp. 125–128. IEEE, 1993.

Cheng-I Lai. Contrastive predictive coding based feature for automatic speaker verification. *arXiv preprint arXiv:1904.01575*, 2019.

Max W. Y. Lam, Jun Wang, Rongjie Huang, Dan Su, and Dong Yu. Bilateral denoising diffusion models, 2021a.

Max WY Lam, Jun Wang, Dan Su, and Dong Yu. Effective low-cost time-domain audio separation using globally attentive locally recurrent networks. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 801–808. IEEE, 2021b.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Jiguo Li, Xinfeng Zhang, Chuanmin Jia, Jizheng Xu, Li Zhang, Yue Wang, Siwei Ma, and Wen Gao. Direct speech-to-image translation. *IEEE Journal of Selected Topics in Signal Processing*, 14(3): 517–529, 2020.

Beata Lorincz, Adriana Stan, and Mircea Giurgiu. Speaker verification-derived loss and data augmentation for dnn-based multispeaker speech synthesis. *arXiv preprint arXiv:2106.01789*, 2021.

Alireza Nasiri and Jianjun Hu. Soundclr: Contrastive learning of representations for improved environmental sound classification. *arXiv preprint arXiv:2103.01929*, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Flavio Protasio Ribeiro, Dinei Florencio, Cha Zhang, and Mike Seltzer. CROWDMOS: An approach for crowdsourcing mean opinion score studies. In *ICASSP*. IEEE. Edition: ICASSP.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 49–58, 2016.

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. *arXiv preprint arXiv:2006.04558*, 2020.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pp. 749–752. IEEE, 2001.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29: 2234–2242, 2016.

Pavan Seshadri and Alexander Lerch. Improving music performance assessment with contrastive learning. *arXiv preprint arXiv:2108.01711*, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pp. 4214–4217. IEEE, 2010.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4879–4883. IEEE, 2018.

Xinsheng Wang, Tingting Qiao, Jihua Zhu, Alan Hanjalic, and Odette Scharenborg. Generating images from spoken descriptions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:850–865, 2021.

Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020.

Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*, 2021.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

Zhen Zeng, Jianzong Wang, Ning Cheng, and Jing Xiao. Lvcnet: Efficient condition-dependent modeling network for waveform generation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6054–6058. IEEE, 2021.

A Diffusion Probabilistic Models

Given i.i.d. samples $\{\mathbf{x}_0 \in \mathbb{R}^D\}$ from an unknown data distribution $p_{data}(\mathbf{x}_0)$. In this section, we introduce the theory of diffusion probabilistic model (Ho et al., 2020; Lam et al., 2021a; Song et al., 2020a;b). First, we present diffusion and reverse process given by denoising diffusion probabilistic models (DDPMs), which could be used to learn a model distribution $p_\theta(\mathbf{x}_0)$ that approximates $p_{data}(\mathbf{x}_0)$. Secondly, we introduce the recently proposed denoising diffusion implicit models (DDIMs) for acceleration. Lastly, we apply bilateral denoising diffusion models (BDDMs) and its tighter evidence lower bound (ELBO) for noise scheduling process, which is efficient in noise schedule prediction.

Diffusion process Similar as previous work (Ho et al., 2020; Lam et al., 2021a; Song et al., 2020a), we define the data distribution as $q(\mathbf{x}_0)$. The diffusion process is defined by a fixed Markov chain from data x_0 to the latent variable x_T:

$$q(x_1, \cdots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{6}$$

For a small positive constant β_t, a small Gaussian noise is added from x_t to the distribution of x_{t-1} under the function of $q(x_t|x_{t-1})$.

The whole process gradually converts data x_0 to whitened latents x_T according to the fixed noise schedule $\beta_1, \cdots, \beta_T$.

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{7}$$

Reverse process Unlike the diffusion process, reverse process is to recover samples from Gaussian noises. The reverse process is a Markov chain from x_T to x_0 parameterized by shared θ:

$$p_\theta(x_0, \cdots, x_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{8}$$

where each iteration eliminate the Gaussian noise added in the diffusion process:

$$p(x_t|x_{t-1}) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)^2 I) \tag{9}$$

Further, denoising diffusion implicit models (DDIM) (Song et al., 2020a) formulate a non-Markovian generative process that accelerates the inference while keeping the same training procedure as denoising diffusion probabilistic models:

From $p_\theta(x_{1:T})$, one can generate a sample x_{t-1} from a sample x_{t-1} via:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta^{(t)}(x_t) + \sigma_t\epsilon_t \tag{10}$$

Training For learning the score network θ, we minimize the bound of the negative log likelihood as described above. Consider a fixed noise schedule, efficient training is optimizing a random term of t with stochastic gradient descent:

$$\mathcal{L}_t = \left\| \boldsymbol{\epsilon}_\theta \left(\alpha_t \mathbf{x}_0 + \sqrt{1-\alpha_t^2}\boldsymbol{\epsilon} \right) - \boldsymbol{\epsilon} \right\|_2^2, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{11}$$

With the noise schedule predictor network, score networks conditioned on discrete-time step and continuous scalar are both possible to take different number of refinement steps for inference (Chen et al., 2020a; Kong et al., 2020b).

Here we define some constants based on the noise schedule in the diffusion process:

$$l_t = \prod_{i=1}^{t} \sqrt{1-\beta_i} \tag{12}$$

$$\alpha_s = l_t, \quad \alpha_{s+1} = l_{t+\tau}, \quad t \sim \text{Uniform}(\{\tau, \dots, T-\tau\}) \tag{13}$$

Where ε denotes the maximum value in a pre-defined linear schedule, and T denotes the total number of diffusion steps.

Noise scheduling acceleration Prior distribution of noise schedule β (e.g., linear) is merely for efficient training at an intermediate step t. Given converged score network θ and noise schedule predictor ϕ, we come to derive a much more efficient noise schedule ζ for revere sampling.

For learning the noise schedule predictor ϕ, we apply the loss function as a KL divergence term between the forward and the reverse distributions

$$\begin{gathered} \mathcal{L}_{\text{step}}^{(t)}(\phi;\theta) = \mathbb{KL}(p_\theta(\mathbf{x}_{t-1}\mathbf{x}_t) \| q_\phi(\mathbf{x}_{t-1}\mathbf{x}_0)) = \\ \frac{1}{2(1-\beta_t-\alpha_t^2)} \| \sqrt{1-\alpha_t^2}\boldsymbol{\epsilon}_t - \frac{\beta_t}{\sqrt{1-\alpha_t^2}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \alpha_t) \|_2^2 + C_t \\ C_t = \frac{1}{4}\log\frac{1-\alpha_t^2}{\beta_t} + \frac{D}{2}(\frac{\beta_t}{1-\alpha_t^2} - 1), \end{gathered} \tag{14}$$

where C_t is a constant that can be ignored during training.

B SCHEDULE ALIGNMENT

Firstly we compute the corresponding constants respective to diffusion and reverse process:

$$l_t = \prod_{i=1}^{t} \sqrt{1-\beta_i}, \quad \alpha_s = \prod_{i=1}^{s} \sqrt{1-\zeta_i} \tag{15}$$

Here we search and interpolate α_s between two training noise constants l_t and l_{t+1}, enforcing α_s to get closed to l_t. In the end, we gain the well-mapped diffusion step t_m:

$$t_m = t + \frac{l_t - \alpha_s}{l_t - l_{t+1}} \quad \text{if } \alpha_s \in [\, l_{t+1}, l_t \,]. \tag{16}$$

Where integer t represents a single pre-defined diffusion step, and s presents a single step of noise schedule obtained through the scheduling process. Given these two reference schedules, schedule alignment could be performed and the floating-point t_m which denotes a much more efficient path for sampling come out.

For noise predictor, we adopt a light-weight GALR network inconsistent with previous work on parameterizing the forward and reverse processes (Lam et al., 2021a). Experimental results further show that the predicted noise schedule ζ for a few samples could be robust enough to maintain a high-quality generation at the reverse process.

C DATASETS

CUB200 CUB dataset (Wah et al., 2011) contains 200 classes and 11788 images in total, which has 8855 and 2933 images for training and testing respectively. For each bird image in CUB dataset, there are 10 speech descriptions.

LJSpeech LJSpeech (Ito & Johnson, 2017) consists of 13,100 short audio clips of a single speaker with a total length of approximately 24 hours.

LibriTTS We use two subsets (i.e., train-clean-360 and train-clean-100) of Multi-speaker English dataset LibriTTS (Zen et al., 2019). The train-clean-360 subset consists of 192 hours' worth of data, 116k utterances, 904 speakers. The train-clean-100 subset consists of audio recordings of 247 speakers with a total duration of about 54 hours.

D EVALUATION MATRIX

LS-MSE and MCD Log-mel spectrogram mean squared error(LS-MSE) and Mel-cepstral distance (MCD) (Kubichek, 1993) measure the consistency between the original waveform and the generated waveform in the Mel-frequency domain.

PESQ and STOI Perceptual evaluation of speech quality (PESQ) (Rix et al., 2001) and The short-time objective intelligibility (STOI) (Taal et al., 2010) assesses the denoising quality for speech enhancement.

MOS All our Mean Opinion Score (MOS) tests are crowdsourced and conducted by native speakers. We refer to the rubric for MOS scores in Protasio Ribeiro et al., and the scoring criteria has been included in Table D for completeness. The samples are presented and rated one at a time by the testers.

Rating	Naturalness	Definition
1	Bad	Very annoying and objectionable dist.
2	Poor	Annoying but not objectionable dist.
3	Fair	Perceptible and slightly annoying dist
4	Good	Just perceptible but not annoying dist.
5	Excellent	Imperceptible distortions

Table 5: Ratings that have been used in evaluation of speech naturalness of synthetic and ground truth samples.

FID FID (Salimans et al., 2016) computes the Fréchet Inception Distance between the distributions of real and generated images in the feature space of pre-trained Inception-V3 network. A smaller FID indicates the synthetic data is more realistic and similar to the true data. We use speech labels in the testing set to generate 10k images.

IS IS (Heusel et al., 2017) is a metric for both image quality and diversity, which is found to correlate well with the human evaluation. In general, a larger IS indicates the generative model can synthesize fake images with better diversity.

Cosine Similarity Cosine Similarity measures the speaker similarity of the synthetic output for the natural samples. Embeddings of utterances from the same speaker have high cosine similarity, while those from different speakers are far apart in the embedding space.

E MULTI-VIEW DATA SYNTHESIS

E.1 MODEL CONFIGURATION

In synthetic data manipulation models, we mainly include variance through duration prediction, energy prediction, and speaker modification, in addition to the diverse input phoneme sequence.

- For content manipulation, we introduce multiple text guidances and keep consistent of remaining features.
- For prosody manipulation, we set the pitch and energy jittering from 0.8 to 1.5, as well as the duration jittering from 0.4 to 1.2.
- For speaker manipulation, we randomly sample speaker identity from train-clean-100 subset of LibriTTS in terms of speaker ID, $\forall \text{ID} \in \{1, \ldots, 850\}$.

FastSpeech consists of 4 feed-forward Transformer blocks both in the encoder and the mel-spectrogram decoder. The hidden sizes of the self-attention and 1D convolution in each feed-forward Transformer block are all set to 256. The number of attention heads is set to 2. The output linear layer converts the 256-dimensional hidden into the 80-dimensional mel spectrogram. In the duration/energy/pitch predictor, the kernel sizes of the 1D-convolution are set to 3.

SynGrad is composed of three location-variable convolution blocks, where each block contains 4 layers, and the residual channels is set to 8. The kernel size of the location-variable convolution is set to three, and the dilation coefficient is the factorial of 3 in each LVCNet block. The kernel predictor is designated to be the same as that in Zeng et al. (2021), where the hidden residual channel is set to 64. The weight normalization is applied in all convolutional layers.

GALR is a speech separation neural network for noise prediction. Following the configuration from previous work (Lam et al., 2021b), we used a window length of 8 samples for encoding, a segment size of 64 for segmentation and only two GALR blocks of 128 hidden dimensions.

E.2 Training and Inference

Raw waveform could be transformed into mel-spectrograms, we set frame size and hop size to 1024 and 256 with respect to the sample rate 22050. Models including FastSpeech 2, SynGrad and GALR have been trained for 15K, 1M and 10K steps until convergence, respectively, using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$.

E.3 Ablation study

To demonstrate the validity of the proposed multi-view data synthesis configuration, we performed an ablation study of several vital parts (i.e., Location variable convolution in 3.1, Noise scheduling acceleration in 3.1) in neural vocoding.

The results of both subjective and objective evaluations are presented in Table 6, which show all three novel designs contribute to the performance: 1) Removing location-variable convolution causes a distinct degradation in generation speed and perceptual quality. 2) Replacing noise prediction with grid search to derive a noise schedule could result in less efficient denoising.

To verify the effect of noise schedule in the settings of diffusion probabilistic models, we compare two noise schedules conditioned on continuous noise levels and discrete noise indices respectively, which are both obtained through noise prediction. We observe that under discrete diffusion indexes, the synthesized samples are cleaner with less noise and the quality improves more stably. The results demonstrate that training SynGrad with discrete noise indexes and applying schedule alignment B with predicted noise schedule could be a better choice.

Model	MOS (↑)	RTF (↓)	LS-MSE (↓)	MCD (↓)	STOI(↑)	PESQ (↑)
Ground Truth	4.65±0.04	/	/	/	/	/
w/o Location variable convolution	4.08±0.05	0.081	140.8	2.3087	0.9719	3.3751
w/o Noise scheduling acceleration	3.95±0.01	0.0328	139.96	2.5832	0.9480	3.0953
Continuous level, 6 steps	3.98±0.08	**0.0296**	108.06	2.4957	0.9709	3.1773
Continuous level, 1000 steps	4.04±0.09	2.80	113.03	2.4899	0.9748	3.2448
Discrete index, 6 steps	4.24±0.09	0.0302	97.49	1.9662	**0.9775**	3.5277
Discrete index, 1000 steps	**4.28±0.09**	4.8185	**81.69**	**1.9242**	0.9669	**3.6525**

Table 6: Ablation study results of several vital components in SynGrad.

F SPEAKER VERIFICATION

F.1 MODEL CONFIGURATION

We train the speaker verification model over train-clean-100 subset of multi-speaker English dataset LibriTTS. The in-domain testing set consists of utterances in test-clean subset, besides we prepare VoxCeleb1 test set following previous work (Fan et al., 2020) to further evaluate the model generalization to out-of-domain distribution. All the models are trained until 100K steps.

The raw waveforms have been sampled to 16000Hz and transformed into mel-spectrograms with a window size of 0.025s and window stride of 0.01s. The speech encoder takes mel-spectrogram as input and outputs an utterance-level fixed-dimensional embedding, which is averaged on time-domain.

- To implement d-vectors models for speaker verification, we follow the configuration in the original paper (Wan et al., 2018).
- For InfoNCE objective function, we explicitly define utterances from same and different speakers to be positive and negative samples, respectively.
- For NT-Xent objective function, we randomly sample with a minibatch of N and define the contrastive prediction task on pairs of utterances from same speaker. Unlike InfoNCE, we would not sample negative examples explicitly but treat the other $2(N-1)$ augmented examples within a minibatch.

G SPEECH-TO-IMAGE TRANSLATION

G.1 MODEL CONFIGURATION

In additional to the the proposed contrastive objective, learning speech representation in speech-to-image translation needs further loss functions for audio-image distillation. DirectGAN adopts Inception-v3 pre-trained on ImageNet as the teacher image encoder, and the speech encoder would be optimized to learn a similar feature space via teacher-student learning. S2IGAN includes joint training of speech and image encoders, and speech representations could be optimized with the supervision of corresponding visual information from images. Model configuration follows the original papers.

The splitting manner for the training set and testing set follows (Reed et al., 2016). Note that we prepare two testing sets for in-domain and out-of-domain evaluation: 1) The seen testing set is made up of utterances from seen speakers in pre-defined prosody ranges as described in E.1. 2) The unseen testing set consists of speakers who do not appear in any training sets, and these speeches are variant in speaking styles (i.e., random deviation of duration, pitch, and energy).

G.2 SAMPLES

In this section, we present in-domain samples in addition to the out-of-domain evaluation of main paper in Figure 4.

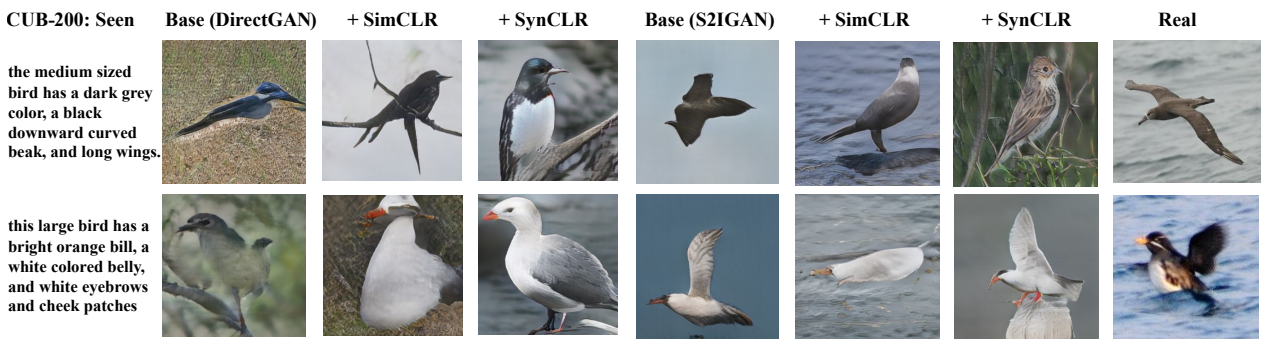


Figure 4: Comparison of synthesized images over in-domain CUB dataset.