S2CD: Self-heuristic Speaker Content Disentanglement for Any-to-Any Voice Conversion

Pengfei Wei¹, Xiang Yin¹, Chunfeng Wang¹, Zhonghao Li¹, Xinghua Qu¹, Zhiqiang Xu², Zejun Ma¹
¹Bytedance, Singapore
²Machine Learning Department, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
¹{pengfei.wei, yinxiang.stephen, wangchunfeng, lizhonghao.01, xinghua.qu, mazejun}01@bytedance.com, ²zhiqiang.xu@mbzuai.ac.ae

Abstract
In this paper, we propose a Self-heuristic Speaker Content Disentanglement (S2CD) model for any-to-many voice conversion with disentanglement without using any external resources, e.g., speaker labels or vectors, linguistic models, and transcriptions. S2CD is built on the disentanglement sequential variational autoencoder (DSVAE), but improves DSVAE structure at the model architecture level from three perspectives. Specifically, we develop different structures for speaker and content encoders based on their underlying static/dynamic property. We further propose a generative graph, modelled by S2CD, so as to make S2CD well mimic the multi-speaker speech generation process. Finally, we propose a self-heuristic way to introduce bias to the prior modelling. Extensive empirical evaluations show the effectiveness of S2CD for any-to-any voice conversion.

Index Terms: voice conversion, any-to-any, disentanglement

1. Introduction
Voice conversion (VC) aims to convert the timbre of a speech from a source speaker to a target speaker while preserving the content of the source speech. Based on the use of training data, VC models can be categorized into parallel and non-parallel ones, where the former is technically simpler but less practical when the target is many speakers and attractive. Precisely, non-parallel VC can be further divided following a “src_to_tgt” naming convention, where src and tgr belong to {one, many, any}. One represents an extreme case where a single speaker is used either in training or inference. In contrast, many and any involve multiple speakers, and the difference is that speakers are seen in many but unseen in any, during the training.

In the last decade, numerous methods have been proposed for non-parallel VC. Although these methods generally share the same idea of learning speaker and content representations, the technical details (e.g., whether using automatic speech recognition (ASR) guided techniques [1,2,3], or text-to-speech (TTS) guided techniques [4, 5, 6], or mixed techniques [7, 8, 9]) and external dependencies (whether using speaker labels/vectors [10, 11, 12], transcriptions [13, 14, 15], or other auxiliary models [16, 17, 18]) differ considerably under different “src_to_tgt” VC scenarios.

Phonetic post-grammars and pre-trained speaker models were widely used for the content and speaker representations in any_to_one VC where the speech corpus of a target speaker is fixed, e.g., [19, 20]. Recently, discretized self-supervised speech representations are proposed to boost any_to_one VC. For instance, Huang et al. [8] propose to use QW2V [21] to eliminate speaker information and represent speech by discrete speech units. SoftVC [7] achieves further improvements by using Hubert [22] to extract soft speech units. Moreover, there are also methods utilizing off-the-shelf TTS techniques to extract speaker-independent linguistic features, e.g., Cotatron [13] takes advantage of Tacotron2 [23] and Mix-Guided VC [9] combines ASR and TTS encoders. However, all these models usually require sufficient training data with rich transcriptions and speaker labels, e.g., [9, 13], and even a large amount of external well-annotated data, e.g., [7, 8, 19], which are very expensive to collect.

Many_to_many VC is a more flexible setting that converts voice among speakers within a speaker training set. Mainstream related research contains generative adversarial network (GAN) based [24, 25, 26, 27] and (variational) autoencoder ((V)AE) based [16, 28, 29, 30]. The key idea of GAN-based VC methods is to learn speaker-indistinguishable representation through adversarial learning. An early work CycleGAN-VC [24] utilizes adversarial and the cycle-consistency losses, but is limited to one-to-one mapping. StarGAN-VC [25] improves to many-to-many mapping by adding another domain classification loss. CC-GAN [26] further uses a speaker-conditional encoder and a multi-output discriminator to simplify the model structure and boost the VC performance. Recently, Ma et al. [27] develop an SGAN-VC using subband block to perform style transfer for each frequency. Note that GAN-based VC methods usually require speaker labels/vectors to train the discriminator. Moreover, the learning objective of these methods usually consists of several losses, e.g., up to 7 losses in [27]. Balancing such many losses is challenging, and the generalization is thus limited.

AE/VAE-based method is another popular research line for many_to_many VC. Basically, these methods aim to disentangle the speaker and content information from speech data. The very first work [28] simply uses a conventional VAE to disentangle the content embedding, and then incorporate it with a pretrained speaker vector for VC. A later work ACVAE-VC [16] uses a speaker-conditioned content posterior and introduces an auxiliary classifier for speaker prediction. Instead of conditioning on speaker attributes, [29] exploits a pitch tracker to constraint an F0-conditioned AE. Note that all these methods need an auxiliary speaker model outside VAE/VAE structure. Recently, Luong et al. [30] propose a disentanglement VAE based on a directed graph that directly models speaker and content latent factors, avoiding the need for external speaker vectors.

Compared with Many_to_many VC, any_to_any VC is a more general scenario, where VC happens between any speakers even they are unseen during training. Due to its generalizability, any_to_any VC research becomes increasingly popular. Pioneering works, AutoVC [10] and AdaIn-VC [31], follow an AE structure and use information bottleneck to separate speaker and content information. A later work VQVC [32] uses vector quantization (VQ) to extract discrete linguistic representations and eliminate the speaker information. However, these meth-
VQMIVC [33] and IDE-VC [34] improve VQVC-1 and AutoVC, respectively, by explicitly building speaker encoder and adding a mutual information (MI) loss, while AGAIN-VC [35] improves AdaIN-VC through a unified encoder and an activation to guide the training. Although technically sound, VQMIVC and IDE-VC suffer from the complex training process due to the difficulty in estimating MI, and AGAIN-VC is not robust to balance the quality of speech audio and the similarity of speaker style.

Except for pretrained speaker models, some any_to_any VC works also take advantage of other external models. Both [36] and [37] adopt wav2vec [38] to extract linguistic embedding, while [39] utilizes speaker verification model [17] to facilitate the speaker modelling. The availability and quality of external models play an important role in the VC success of these methods. GAN-based ideas are also exploited in any_to_any VC, e.g., [11, 12, 40]. However, same as stated in many_to_many VC, these methods usually stack a large amount of losses, specifically 5 in [40, 11] and 7 in [12], and thus lack of generalizability. There are also methods achieving VC by TTS-based system, e.g., YourTTS [6] and STYLETTS [15], with rich transcription or phonemes available for training.

Very recently, VAE-based methods have shown a great success in any_to_any VC. In [41], the authors propose a variant of β-VAE [42] that is specifically for disentanglement of content and speaker representations by two individual latent factors. A later work [43] further investigates a more powerful VAE model, disentangled sequential VAE (DSVAE) [44], specifically disentangling time-invariant and time-variant information from sequential data. Such VAE-based methods are elegant for any_to_any VC in the sense that they fully rely on the strong disentanglement capability of model itself without using any external resources (e.g., pretrained linguistic models, transcriptions and speaker labels/vectors) and auxiliary losses (only reconstruction and KL losses are used in the learning objective).

A very recent work CDSVAE [14] further shows that the VC performance can be boosted using external models or transcriptions to introduce content bias to the prior modeling.

In this paper, we aim to propose a novel VC method without using any external resources for any_to_any VC. We build on DSVAE that explicitly models speaker and content latent factors. However, different from [14, 43] that directly adopt the original structure of DSVAE [44], we develop more advanced submodule structures at the model architecture level to better serve for VC purpose, meanwhile, without adding extra losses to vanilla DSVAE objectives. Specifically, we propose the following three improvements.

Firstly, we design different structures for content and speaker encoders, rather than to use the same one, i.e., BiLSTM, as [14, 43, 44] do. Considering that content and speaker latent factors encode dynamic and static information, respectively, we propose to use BiLSTM and transformer without positional embedding as the base model of content and speaker encoders, correspondingly. We intuitively and empirically show that such design benefits the disentanglement\(^1\).

Secondly, to further enhance the benefit of disentanglement to VC, we enforce our model to follow a generative graph as shown in Fig.1. Ideally, identical utterance of different speakers is generated from the same content latent factor but different speaker latent factors, while different utterances of the same speaker should share the same speaker latent factor. As parallel data is absent, we put more focus on the latter. Specifically, we propose to feed positive pairs of utterances (utterances from the same speaker) into speaker encoder to model a shared speaker latent factor by using an average function.

Thirdly, we also introduce related bias to the prior modelling. However, instead of using external speaker or linguistic models, we propose a self-heuristic way. We build the prior directly using speaker and content representations sampled from the corresponding posterior. With these improvements, we obtain our Self-heuristic Speaker Content Disentanglement model (S2CD) for any_to_any VC. To summarize, the main contribution of this paper is as follows.

- We present a comprehensive review of existing non-parallel VC methods, discussing their technical details and external dependencies under different “src_totgt” VC scenarios.
- We propose an S2CD model for any_to_any VC without using any external resources. S2CD-VC is based on DSVAE, but improves DSVAE from three perspectives.
- We conduct extensive empirical evaluations, including comparison with existing methods and property analyses, on S2CD to show its effectiveness for any_to_any VC.

2. The Proposed Method

2.1 Preliminary

We start with the problem formulation of any_to_any VC. Let \(X = [x_1, ..., x_T]\) be a \(T\)-segment speech, represented by acoustic features, e.g., mel-spectrogram, of a speaker sampled from \(S = [s_1, ..., s_n]\). Our goal is to train a model with the speech data of multiple speakers from \(S\) without using any external resources, for any_to_any VC. The test includes two scenarios, namely seen2seen: \(s_i \rightarrow s_j\) where \(s_i, s_j \in S\) and unseen2unseen: \(s_i \rightarrow s_j\) where \(s_i, s_j \notin S\).

As our S2CD uses DSVAE as the backbone, we introduce DSVAE herein. The latest DSVAE architecture adopted in [14] is shown in Fig.2(a). The input mel-spectrograms are fed into a shared encoder, which consists of several convolutional layers, to extract the deep acoustic features, followed by a BiLSTM to explore the temporal information. Two groups of mean and variance networks are applied to model the posterior of speaker and content latent factors, i.e., \(q_\theta(z|x, \lambda)\) and \(q_\theta(z|x, \lambda)\). The new representations \(z_1^* …, z_T^*\) and \(z^*\) are correspondingly sampled. Each \(z_i^*\) is then concatenated with \(z^*\) and passed to a decoder for reconstruction. Finally, the reconstructed mel-spectrograms are gone through a vocoder to construct the waveform. The prior of \(z^*\) and \(z_i^*\) is a standard Gaussian distribution and modelled by an autoregressive LSTM, respectively. Both posterior and prior distributions follow the conditional independence assumption same as [44]. The overall learning objective then consists of reconstruction and KL-divergence parts:

\[
\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{kl-c}} + \lambda_2 \mathcal{L}_{\text{kl-c}},
\]

\(^1\)Using different structures for encoder is a general improvement for DSVAE [44] beyond VC task, as it essentially aims to strengthen the disentanglement of static and dynamic information.
3. Experiments

3.1. Experimental Configuration

We use VCTK dataset [47] for experimental study. Following [14], we randomly select 10 speakers forming unseen speaker
Figure 3: Visualizations of speaker latent embedding in the unseen2unseen scenario.

Table 1: The MOS test with 95% CI.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Seen2Seen</th>
<th>Unseen2Unseen</th>
</tr>
</thead>
<tbody>
<tr>
<td>AutoVC</td>
<td>1.37 ±0.06</td>
<td>2.85 ±0.24</td>
</tr>
<tr>
<td>DSVAE</td>
<td>3.18 ±0.09</td>
<td>2.93 ±0.18</td>
</tr>
<tr>
<td>CDSVAE [14]</td>
<td>3.84 ±0.10</td>
<td>3.31 ±0.18</td>
</tr>
<tr>
<td>DSVAE</td>
<td>3.76 ±0.07</td>
<td>3.83 ±0.06</td>
</tr>
<tr>
<td>CDSVAE [14]</td>
<td>4.03 ±0.04</td>
<td>4.12 ±0.01</td>
</tr>
<tr>
<td>S2CD w/o T</td>
<td>4.09 ±0.09</td>
<td>3.91 ±0.13</td>
</tr>
<tr>
<td>S2CD</td>
<td>4.11 ±0.06</td>
<td>4.02 ±0.09</td>
</tr>
<tr>
<td>S2CD</td>
<td>4.32 ±0.04</td>
<td>4.21 ±0.07</td>
</tr>
</tbody>
</table>

Table 2: Phoneme classification accuracy (%).

<table>
<thead>
<tr>
<th>Methods</th>
<th>Mel-only</th>
<th>DSVAE</th>
<th>S2CD w/o T &amp; P</th>
<th>S2CD w/o P</th>
<th>S2CD w/o T</th>
<th>S2CD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>59.75</td>
<td>47.2</td>
<td>63.95</td>
<td>64.47</td>
<td>64.77</td>
<td></td>
</tr>
</tbody>
</table>

set, while the rest speakers are used for training. We extract melspectrogram as features with a framing configuration of 64ms/16ms, and set feature dimension to 80. We select a segment of 64 frames for training. Adam is used as the optimizer with fixed learning rate of 1e-4. The dimension of the convolutional layers in shared encoder is 256, and that of speaker and content latent factors is set to 64. Batch size is set to 256. For \( \lambda_s \) and \( \lambda_c \), we use 0.1 and 1, respectively, according to the grid search. All the experiments are done on NVIDIA A100 GPU.

3.2. Subjective Evaluation

We first evaluate S2CD by the mean opinion score (MOS) test. For both seen2seen and unseen2unseen scenarios, we randomly select 30 test cases. Each test case includes two utterances from a source speaker and a target speaker randomly selected from the corresponding speaker set, as well as the converted utterance. Ten listeners evaluate the converted utterances by giving scores from 1 to 5 on naturalness and similarity. The final score is calculated by averaging all the collected results. The demo samples are shown in this link\(^2\).

We compare with several baselines including AutoVC [10], DSVAE and CDSVAE [14]. As the code of [14] is not public, we implement it according to the paper. Unfortunately, we fail to reproduce the scores\(^3\). Thus we show the results of our run and also the paper-reported numbers. For S2CD, we introduce two more variants, namely S2CD w/o T & P and S2CD w/o T. We gradually add the three improvements to DSVAE, firstly the pair-wise training, resulting in S2CD w/o T & P, followed by self-heuristic prior modelling, resulting in S2CD w/o T, and lastly customized encoders, giving us the final S2CD.

The MOS results are shown in Fig.1. As can be seen, our reproduced DSVAE and CDSVAE results are generally lower than the paper-reported ones, but the trend is the same, i.e., CDSVAE improves DSVAE. We mainly compare with the reported scores. For our methods, S2CD w/o T & P achieves comparable and better results compared with DSVAE on similarity and naturalness, respectively, but is still not as good as CDSVAE, especially on similarity. By adding self-heuristic prior modelling, S2CD w/o T generally catches up with CDSVAE with better naturalness but weaker similarity. Further taken customized encoders into account, S2CD finally outperforms CDSVAE in average. Note that S2CD has another superiority over CDSVAE, i.e., its independence on external resources. The performance gain from S2CD w/o T & P to S2CD also shows the effectiveness of each improvement for VC.

3.3. Latent Factors Analyses

We also analyze the disentanglement performance by (1) visualizing speaker latent embeddings and (2) performing phoneme classification with content latent embeddings. We show the t-SNE visualization and phoneme classification results on the test speaker set in Fig.3 and Table 2, respectively. DSVAE obtains a clear cluster pattern in Fig.3(a). However, the worst phoneme classification accuracy of DSVAE, 47.3% even worse than mel_only, is also observed. This shows the room of further improvements on disentanglement. For S2CD w/o T & P, it has looser clusters but better phoneme classification performance than DSVAE, which overall balances the performance gain. This is consistent with their comparable MOS in Table 1. S2CD w/o T achieves similar distributed clusters but clear better accuracy than DSVAE. Thus, we observe the performance gain of S2CD w/o T over DSVAE on MOS. For the final S2CD, it achieves not only denser clusters but also higher accuracy compared with all the other baselines, and thus is the best on VC performance. This set of experiments show that our proposed improvements indeed lead to a better disentanglement for VC.

4. Conclusions

In this paper, we first present a comprehensive review of existing non-parallel VC methods under different "any-to-any" scenarios. We then focus on the latest "any-to-any" model DSVAE, and propose an S2CD model with three improvements, i.e., customized encoder structures, positive pair-wise training, and self-heuristic prior modelling, over DSVAE. Empirical results show S2CD is a promising method for "any-to-any" VC.

---

\(^2\)https://wmaiga.github.io/S2CD/

\(^3\)The model structure is well stated in [14], and we suspect the non-reproducibility is due to some imperceptible differences of our implementation from the official one.
5. References


