CNEG-VC: Contrastive Learning using Hard Negative Example in Non-parallel Voice Conversion

Bima Prihasto, Yi-Xing Lin, Phuong Thi Le, Chien-Lin Huang, and Jia-Ching Wang

Contact Information:
Computer Science and Information Engineering
National Central University
Zhongli District, Taoyuan City, Taiwan

Email: `jcw@csie.ncu.edu.tw`

Abstract

Contrastive learning has advantages for non-parallel voice conversion, but the previous conversion results could be better and more preserved. In previous techniques, negative samples were randomly selected in the features vector from different locations. A positive example could not be effectively pushed toward the query examples. We present contrastive learning in non-parallel voice conversion to solve this problem using hard negative examples. We named it CNEG-VC. Specifically, we teach the generator to generate negative examples. Our proposed generator has specific features. First, Instance-wise negative examples are generated based on voice input. Second, when taught with an adversarial loss, it can produce hard negative examples. The generator significantly improves non-parallel voice conversion performance. Our CNEG-VC achieved state-of-the-art results by outperforming previous techniques.

Introduction

Voice conversion involves transforming the speaker's voice from the source speaker towards the target speaker while keeping the information in the content. Recent methods are typically based on a non-parallel setting due to the inconvenience of collecting parallel training data. The generative adversarial network (GAN) can get satisfactory results with this method. Where is the cycle-consistency loss was extensively used to maintain the consistency of generated, and source speech, such as StarGAN-VC [3] and CycleGAN-VC3 [1]. Remarkably, the recently proposed framework CVC [2] based on contrastive learning with the cycle-consistency loss improves performance over frameworks [1].

However, the negative samples within the framework [2] are drawn randomly from various locations within the mel-spectrogram domain. Consequently, the converted results are sometimes of poor quality, and the content could be more consistently preserved. In other words, these negative samples are insufficient to bring the positive samples closer to the query samples, trying to prevent the framework from optimizing the advantages of contrastive learning.

Contribution

- We present contrastive learning in non-parallel voice conversion to solve the current issue using hard negative examples (CNEG-VC).
- We present a new negative generator developed to reveal hard negative examples. The negative generator generates instance-wise negative examples based on the embedded feature.
- We add noise to the generator as an additional input. The generator can neglect the noise input, producing similar examples for different noise inputs.
- We introduce a diversity loss in the generator to encourage the generator to generate different hard negative examples for various input noises.

Methods

As in Fig1, the three parts of our framework are the speech generator encoder, the negative generator, and the representation network. The speech generator encoder generates positive and negative samples on a particular layer. The negative generator works to mine the instance-wise hard negative example to trigger the positive samples to increase the correlation between the positive and the query samples. High-dimensional space of the representation network contains spatially embedded feature vectors.

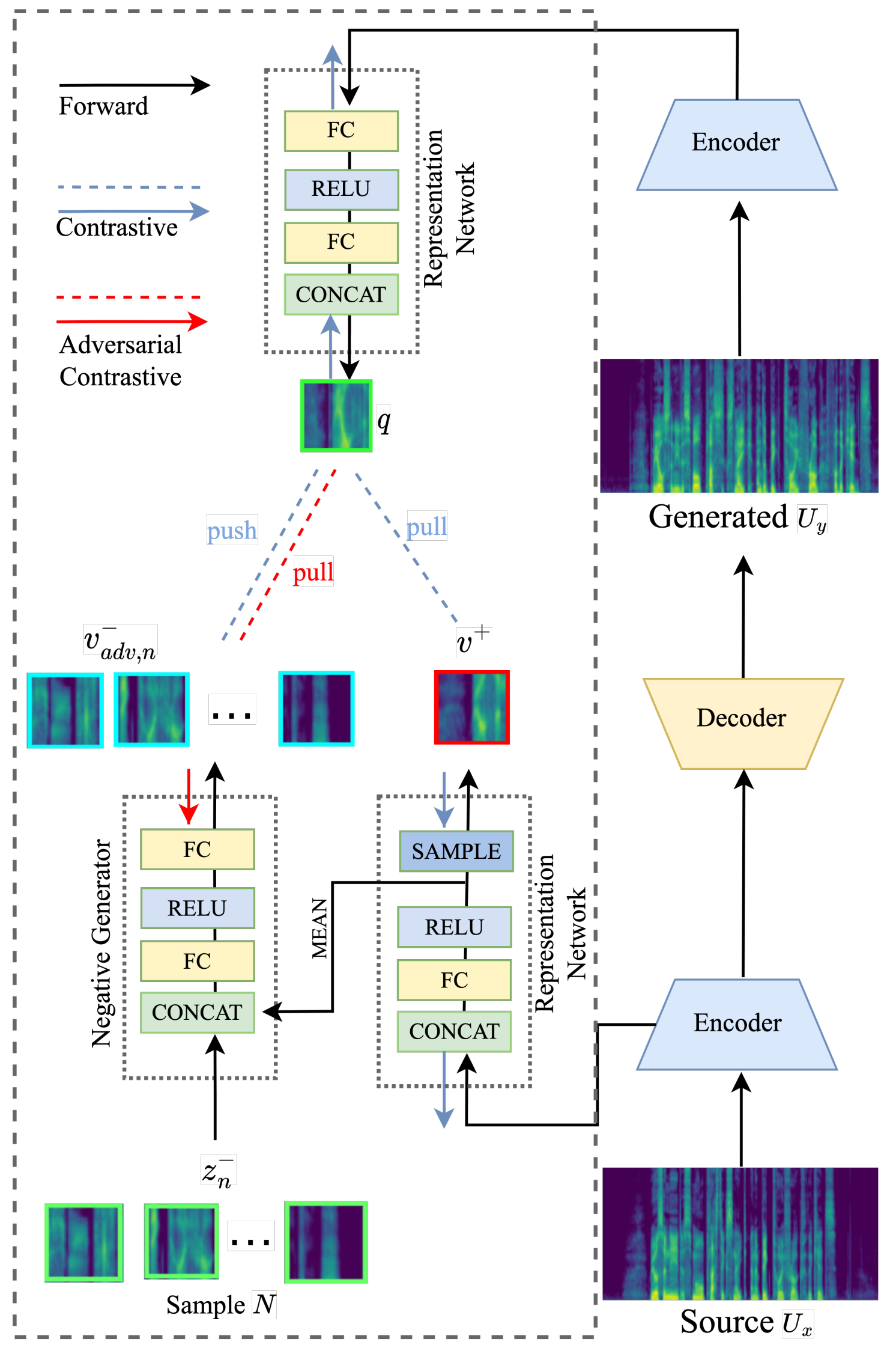


Figure 1: The overview of our CNEG-VC framework

Hard Negative Example Method

We employ the representation network $R^i(\cdot)$ used to embed feature vectors on a particular layer in the encoder network. In Eq.1, The speech generator encoder defines the source and generated features vectors as $F_i^{U_x}$ and $F_i^{U_y}$ with $i-$th layers.

$$q = \frac{R_s^i(F_i^{U_y})}{\left\| R_s^i(F_i^{U_y}) \right\|_2}, v^+ = \frac{R_s^i(F_i^{U_x})}{\left\| R_s^i(F_i^{U_x}) \right\|_2} \tag{1}$$

In Eq.2, we added a noise vector z_n to the negative generator for generating hard negative examples.

$$v^-_{adv,n} = \frac{N^i(\overline{R_s^i(F_i^{U_x})}; z_n)}{\left\| N^i(\overline{R_s^i(F_i^{U_x})}; z_n) \right\|_2} \tag{2}$$

Contrastive learning aims to train the negative generator adversarially against the encoder network, as shown in Eq.3.

$$\begin{aligned} &\min_{\theta_R,\theta_G} \max_{\theta_N} l(q, v^+, v^-_{adv}) = \\ &- \log \left[\frac{\exp(q \cdot v^+/\tau)}{\exp(q \cdot v^+/\tau) + \sum_{n=1}^{N} \exp(q \cdot v^-_{adv,n}/\tau)} \right] \end{aligned} \tag{3}$$

The process of updating the weights on the negative generator and the negative contrastive is defined by Eq.4.

$$\theta_{N^i} \leftarrow \theta_{N^i} + \eta_{\mathcal{N}} \frac{\partial l(q, v^+, v^-_{adv})}{\partial \theta_{N^i}} \tag{4}$$

After that, the network representation is updated with the positive contrastive loss defined by Eq.5.

$$\theta_{R^i} \leftarrow \theta_{R^i} - \eta_{\mathcal{R}} \frac{\partial l(q, v^+, v^-_{adv})}{\partial \theta_{R^i}} \tag{5}$$

Total adversarial contrastive loss of representation network and negative generator is defined by Eq.6.

$$\mathcal{L}_{AdvC} = \mathbb{E}_{x \sim U_x} \sum_{l=1}^{L} \sum_{s=1}^{S_l} l(q_{l,s}, v^+_{l,s}, v^-_{adv,l,s}) \tag{6}$$

The speech generator encoder is updated with a weighted sum of these gradients, as defined by Eq.7.

$$\theta_G \leftarrow \theta_G - \eta_G \sum_{i=0}^{l} \left(\frac{\partial l(q, v^+, v^-_{adv})}{\partial F_i^{U_x}} \frac{\partial F_i^{U_x}}{\partial \theta_G} + \frac{\partial l(q, v^+, v^-_{adv})}{\partial F_i^{U_x}} \frac{\partial F_i^{U_y}}{\partial \theta_G} \right) \tag{7}$$

The diversity loss emphasizes obtaining reliable results when adding different noise vectors, defined by Eq.8.

$$\mathcal{L}_{dn} = - \left\| N^i(\overline{R^i(U_{x_i})}, z_1) - N^i(\overline{R^i(U_{x_i})}, z_2) \right\|_1 \tag{8}$$

We unitize the LSGAN loss for the speech generator encoder G and discriminator D, defined by Eq.9.

$$\begin{aligned} \mathcal{L}_{gan}^{D} &= E_{x_r \in U_x} \log(1 - D(x_r) + E_{x_f \in U_y} \log(1 - D(x_f) \\ \mathcal{L}_{gan}^{G} &= E_{x_f \in U_y} \log(1 - D(x_f) \end{aligned} \tag{9}$$

Overall, the loss is calculated using a weighted summation derived from the encoder network and a negative generator defined by Eq.10.

$$\begin{aligned} &\mathcal{L}_R = \mathcal{L}_{AdvC}; \mathcal{L}_G = \mathcal{L}_{AdvC} + \lambda_1 \mathcal{L}_{gan}^{G} \\ &\mathcal{L}_N = -\mathcal{L}_{AdvC} + \lambda_2 \mathcal{L}_{dn} \end{aligned} \tag{10}$$

Results

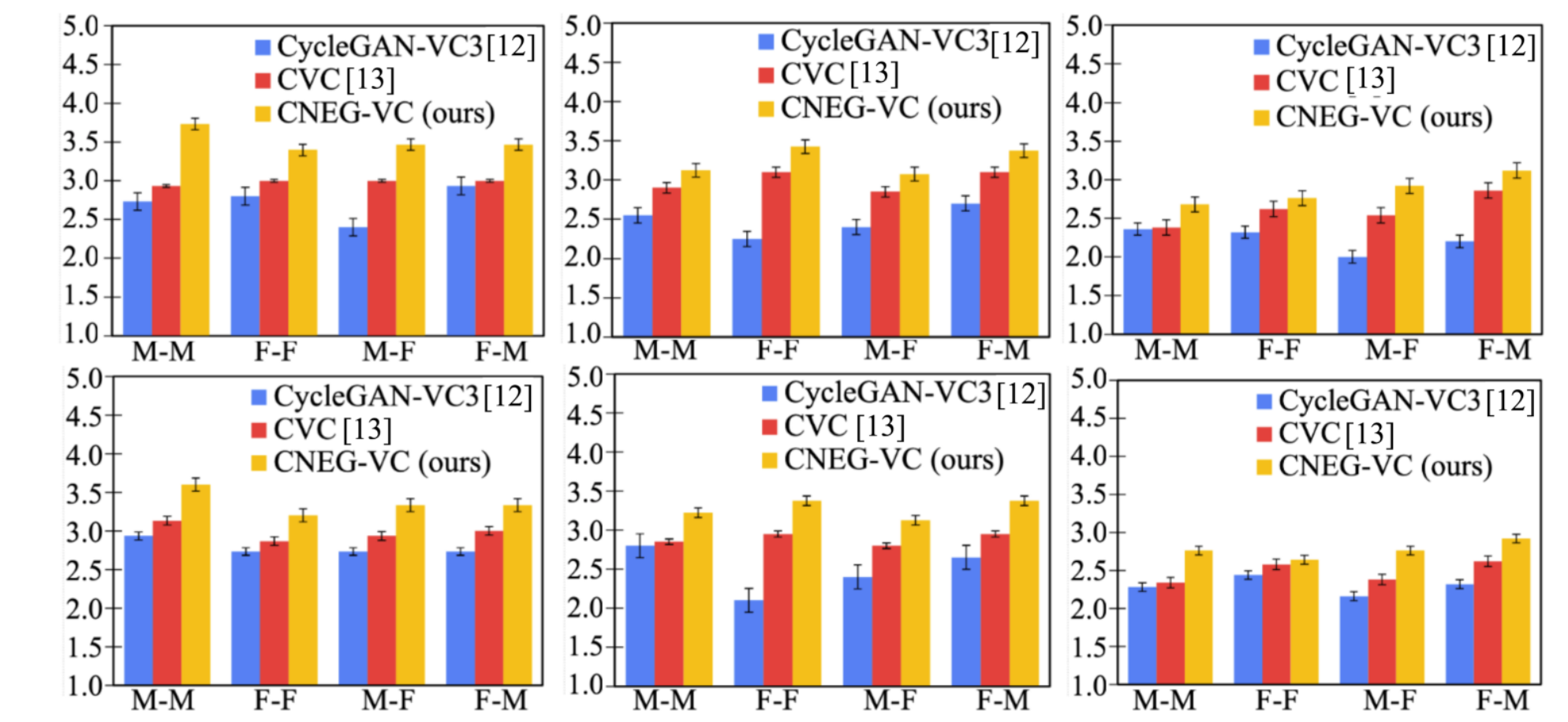
The quantitative results of the objective evaluation are shown in Table 1. Our method outperforms CycleGAN-VC3, proving that contrastive learning can substantially impact. Our method is superior to CVC, which also employs contrastive learning. This demonstrates that these negative examples are sufficient to bring the positive examples closer to the query examples, allowing the framework to take advantage of contrastive learning. In addition to light results, our method has also been evaluated with excellent outcomes. It is interesting to note that the results will be substantially worse if we use fewer training data.

Table 1: Objective evaluation results with voice similarity. The results value indicates [standard data] / [light data] at the training stage.

Scheme	Gender	Voice Similarity		
		CycleGAN-VC3 [1]	CVC [2]	CNEG-VC(ours)
One to one	Male-Male ↑	0.962 / 0.899	0.964 / 0.908	**0.968 / 0.918**
	Female-Female ↑	0.935 / 0.879	0.937 / 0.882	**0.945 / 0.890**
	Male-Female ↑	0.925 / 0.854	0.929 / 0.862	**0.934 / 0.876**
	Female-Male ↑	0.951 / 0.895	0.952 / 0.908	**0.963 / 0.919**
Many to One	Male-Male ↑	0.923 / 0.905	0.934 / 0.907	**0.943 / 0.921**
	Female-Female ↑	0.929 / 0.854	0.935 / 0.869	**0.940 / 0.880**
	Male-Female ↑	0.926 / 0.887	0.937 / 0.892	**0.945 / 0.915**
	Female-Male ↑	0.968 / 0.898	0.974 / 0.921	**0.976 / 0.926**
Many to One (unseen)	Male-Male ↑	0.907 / 0.834	0.913 / 0.851	**0.927 / 0.856**
	Female-Female ↑	0.889 / 0.816	0.911 / 0.833	**0.929 / 0.848**
	Male-Female ↑	0.910 / 0.846	0.925 / 0.855	**0.937 / 0.867**
	Female-Male ↑	0.935 / 0.802	0.945 / 0.810	**0.957 / 0.830**

Our method outperformed the previous method in the subjective evaluation shown in Fig.2. The experimental results indicate that one-to-one is marginally superior to the many-to-one scheme. Compared to speaker similarity, our approach consistently yields high naturalness values. Subjective evaluation reveals if the sourced-from-male voices scheme performs well in a one-to-one setting, the sourced-from-female voices scheme tends to perform better in a many-to-one setting.

In the ablation study, several conditions were utilized to evaluate the effectiveness of our CNEG-VC. Table 2 demonstrates that the results would be worse without negative generators and diversity loss. Negative examples are only sufficiently challenged with negative generators or diversity loss. Results among all gender settings indicate that results will improve as negative generators and diversity are used.



(a) One-to-one (b) Many-to-one (c) Many(unseen)-to-one

Figure 2: Subjective evaluation results with MOS. Naturalness (upper) and speaker similarity (lower), with a 95% confidence interval.

Table 2: Ablation Experiment.

Settings		Gender			
Negative Generator	Diversity Loss	M-M ↑	F-F ↑	M-F ↑	F-M ↑
×	×	0.944	0.921	0.911	0.939
✓	×	0.963	0.939	0.929	0.958
✓	✓	**0.968**	**0.945**	**0.934**	**0.963**

Conclusions

This paper proposes CNEG-CV as a novel framework for non-parallel voice conversion with a hard negative examples approach for contrastive learning. We constructed a negative generator for adversarial supervision against the encoder network. Encoder networks and negative generators are trained to differentiate positive examples from hard negative examples that have been generated through diversity loss. Our CNEG-VC has achieved better results than the previous method and proposed state-of-the-art voice conversion.

References

[1] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-Spectrogram Conversion. In *Proc. Interspeech*, 2020.

[2] Tingle Li, Yichen Liu, Chenxu Hu, and Hang Zhao. CVC: Contrastive Learning for Non-Parallel Voice Conversion. In *Proc. Interspeech*, 2021.

[3] Yinghao Aaron Li, Ali Zare, and Nima Mesgarani. StarGANv2-VC: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion. In *Proc. Interspeech*, page 1349–1353, 2021.