StarGAN-VC2:

Rethinking Conditional Methods for StarGAN-Based Voice Conversion

Takuhiro Kaneko   Hirokazu Kameoka   Kou Tanaka   Nobukatsu Hojo

NTT Communication Science Laboratories, NTT Corporation, Japan

StarGAN-VC2 samples

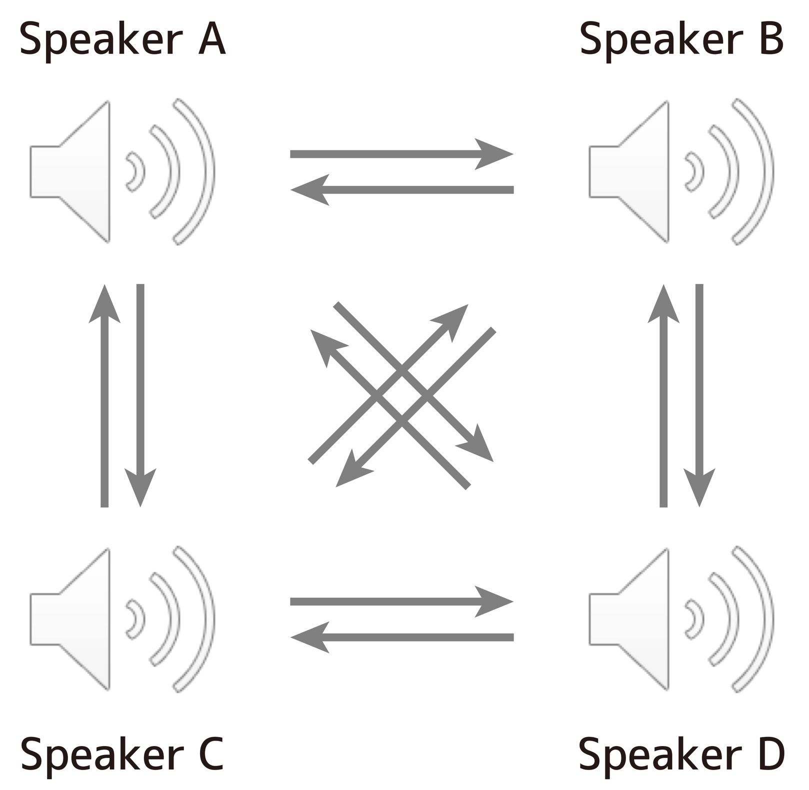
INTERSPEECH 2019

Introduction

Objective

Non-parallel multi-domain VC

- Our goal is to learn mappings among multiple domains (e.g., multiple speakers) without relying on parallel data.

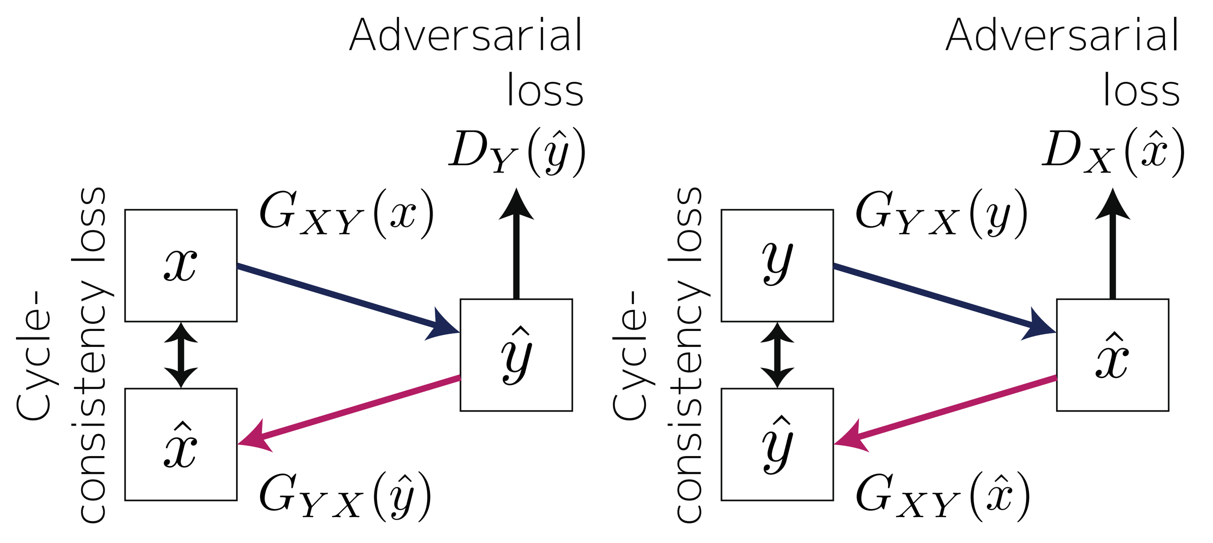


Limitations of one-to-one VC

E.g., CycleGAN-VC [Kaneko+2017]

- Achieves one-to-one VC in a non-parallel setting.
- However, requires **many generators** to achieve multi-domain VC.

(Their number increases according to the number of domains)



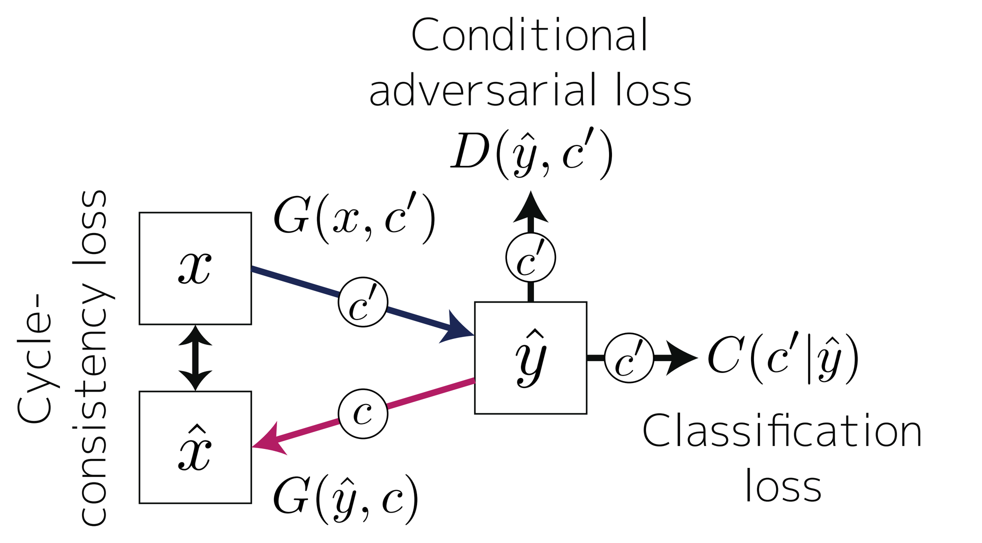
Forward-inverse mapping   Inverse-forward mapping

Possible solution

StarGAN-VC [Kameoka+2018]

- Extends CycleGAN-VC to a **conditional setting** by incorporating domain codes.
- Only requires a **single generator**.
- However, **the quality is still low**.

→ **Challenge to address**



c: Source domain code, c' : Target domain code

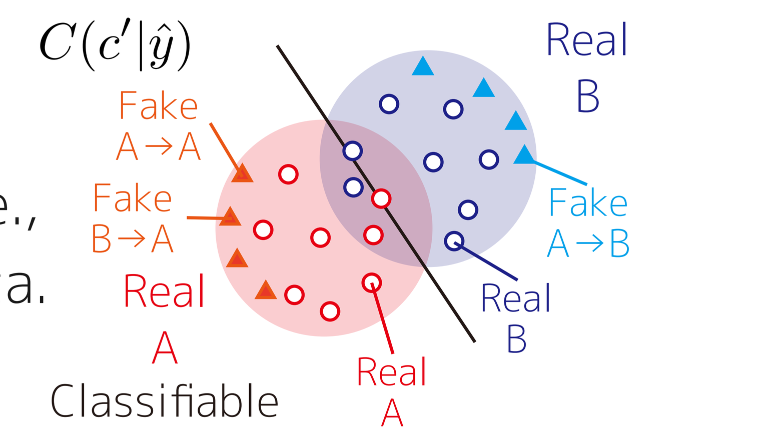
Proposed method: StarGAN-VC2

- **Key ideas:** We rethink **conditional methods** of StarGAN-VC in two aspects: **training objectives** and **network architectures**.

1. Rethinking conditional methods in training objectives

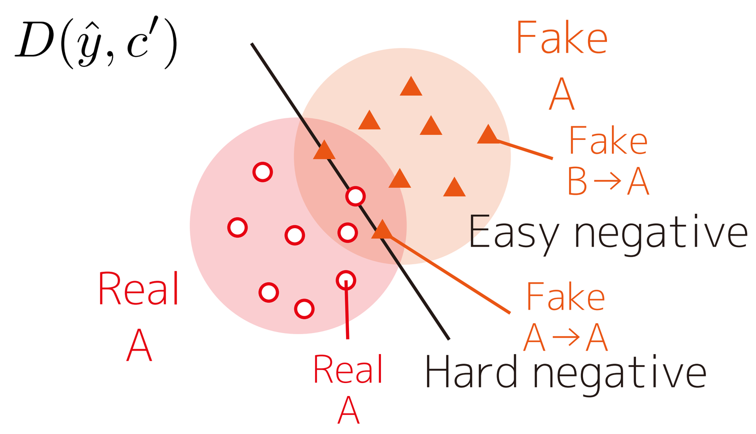
i. (Previous) Classification loss

- C is learned using real data.
- G tries to generate **classifiable** (i.e., far from the decision boundary) data.



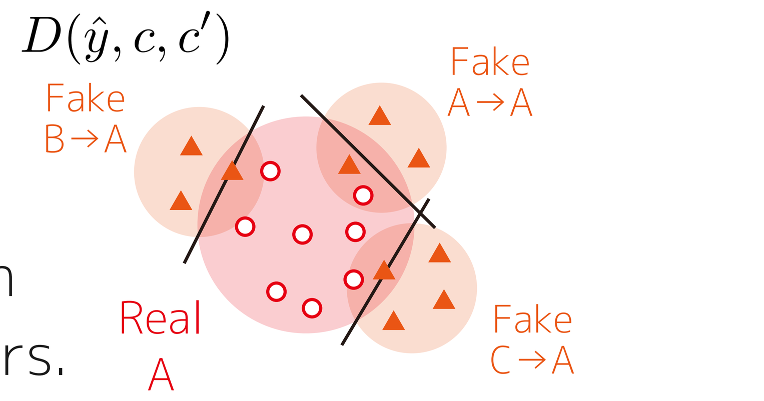
ii. (Previous) Target conditional adversarial loss

- D needs to simultaneously handle **hard negative** (e.g., A→A) and **easy negative** (e.g., B→A) samples.



iii. (Proposed) Source and target conditional adversarial loss

- This loss brings **all the converted data** close to the target data in both **source-wise** and **target-wise** manners.



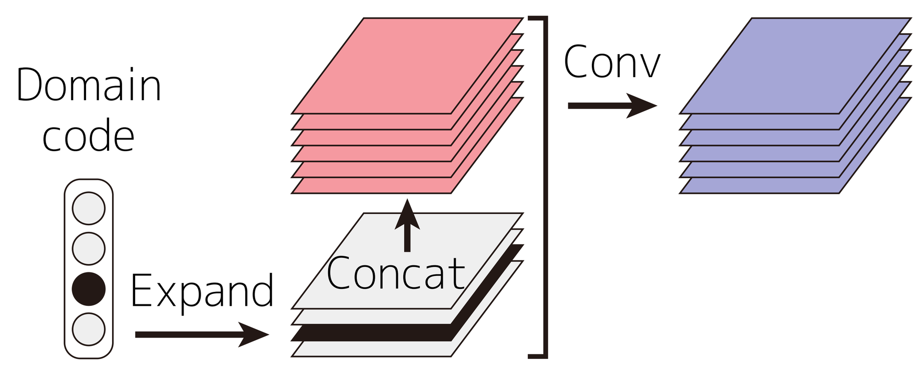
2. Rethinking conditional methods in G networks

i. Motivation

- Accurate **modulation translation** is important to achieve high-quality VC (e.g., GV [Toda+2007] & MS [Takamichi+2014] postfilters).

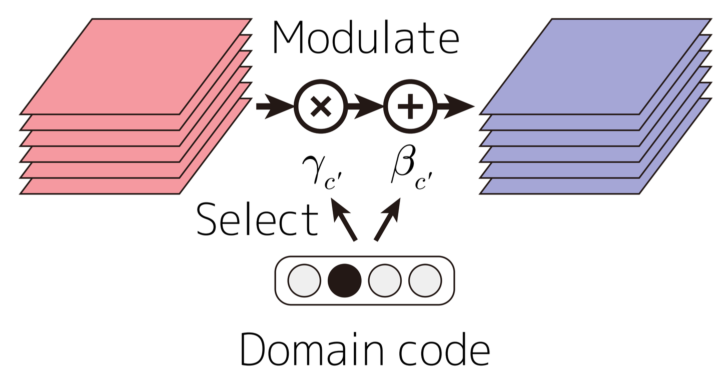
ii. (Previous) Channel-wise

- Concatenated domain codes are **additively** used.
- They **cannot be directly used** for **modulating** data.



iii. (Proposed) Modulation-based

- Domain codes are used to **select modulation parameters**.
- They **can be directly used for modulating** data.



$$\mathrm{CIN}(\boldsymbol{f}; c') = \gamma_{c'} \left(\frac{\boldsymbol{f} - \mu(\boldsymbol{f})}{\sigma(\boldsymbol{f})} \right) + \beta_{c'}$$

Conditional instance normalization [Dumoulin+2017]

Experiments

Experimental conditions

i. Data

- **Dataset:** Voice Conversion Challenge 2018
- **Speakers:** 4 Professional US English speakers (VCC2SF1, VCC2SF2, VCC2SM1, and VCC2SM2)
- **Sentences:** 81 sentences (about 5 min.)
- **Sampling Rate:** 22.05 kHz
- **Features:** 34 MCEPs, log F_0, APs (WORLD, 5 ms)

ii. Conversion process (Follow VCC 2018 baseline)

- **MCEP:** StarGAN-VC2
- **log F_0:** Linear transformation
- **AP:** No conversion

→ WORLD vocoder [Morise+2016]

iii. Implementation and training

- Network architectures are based on **CycleGAN-VC2** [Kaneko+2019] (G: 2-1-2D CNN, D: 2D CNN).
- In training, **no extra data, modules, or time alignment procedure** are used.
- 4 × 3 = 12 different source-and-target mappings are learned in a **single generator**.

Objective evaluation

i. Evaluation metrics

- **Mel-cepstral distortion (MCD):** Global structural difference (smaller is better)
- **Modulation spectra distance (MSD):** Local structural difference (smaller is better)

ii. Comparison of training objectives

Objective	MCD [dB]	MSD [dB]
$\mathcal{L}_{cls}$	7.73 ± .07	1.96 ± .03
$\mathcal{L}_{t\text{-}adv}$	7.21 ± .16	2.87 ± .51
$\mathcal{L}_{t\text{-}adv} + \mathcal{L}_{cls}$ (StarGAN-VC)	7.11 ± .10	2.41 ± .13
$\mathcal{L}_{st\text{-}adv}$ (StarGAN-VC2)	6.90 ± .07	1.89 ± .03

- Improves both MCD and MSD.

Note: We fix the conditional method in G networks as modulation-based.

iii. Comparison of G networks

G network	MCD [dB]	MSD [dB]
Channel-wise (StarGAN-VC)	6.90 ± .08	2.55 ± .20
Modulation-based (StarGAN-VC2)	6.90 ± .07	1.89 ± .03

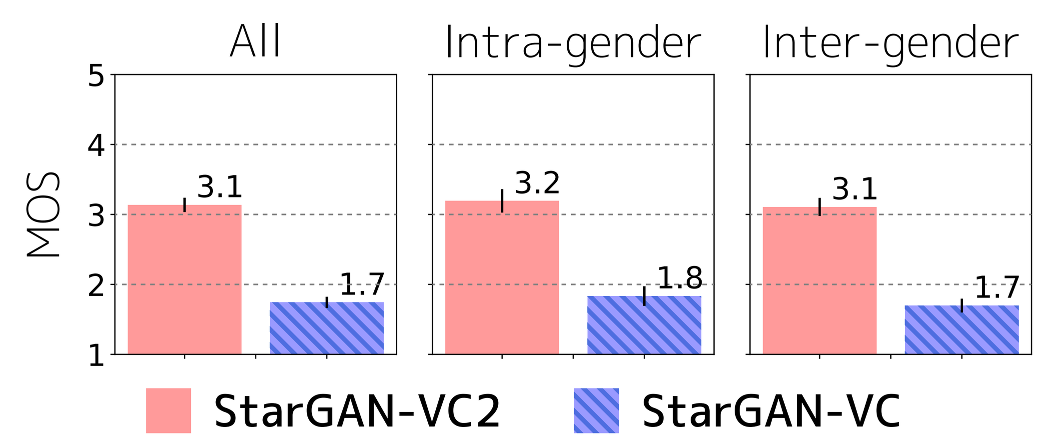
- Improves MSD.

Note: We fix the conditional method in the training objectives as $\mathcal{L}_{st\text{-}adv}$.

Subjective evaluation

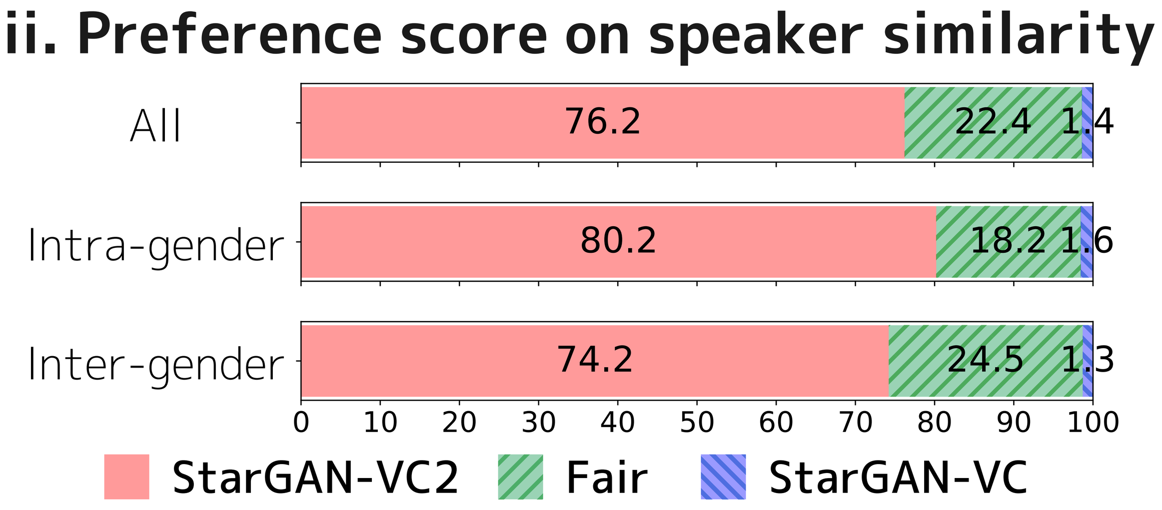
StarGAN-VC [Kameoka+2018] vs. StarGAN-VC2

i. MOS for naturalness



- StarGAN-VC2 outperforms StarGAN-VC for every category.

ii. Preference score on speaker similarity



- StarGAN-VC2 outperforms StarGAN-VC for every category.