Objective

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

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., $A \rightarrow A$) and easy negative (e.g., $B \rightarrow A$) 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.

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**
  - **MCEP:** StarGAN-VC2
  - **log $F_0$:** Linear transformation
  - **AP:** No conversion
  - **WORLD vocoder** [Morise+2016]

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.
- $4 \times 3 = 12$ different source-and-target mappings are learned in a single generator.

Subjective evaluation
- **i. MOS for naturalness**
  - **StarGAN-VC2 [Kameoka+2018] vs. StarGAN-VC**
  - StarGAN-VC2 outperforms StarGAN-VC for every category.
- **ii. Preference score on speaker similarity**
  - All
    - StarGAN-VC2: 76.2, StarGAN-VC: 32.4
  - Intra-gender
    - StarGAN-VC2: 80.2, StarGAN-VC: 18.2
  - Inter-gender
    - StarGAN-VC2: 74.2, StarGAN-VC: 34.5

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**

Copyright © 2019 Nippon Telegraph and Telephone Corporation