CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion

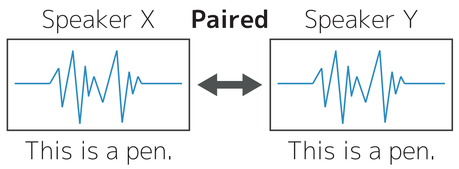
Takuhiro Kaneko Hirokazu Kameoka Kou Tanaka Nobukatsu Hojo
NTT Communication Science Laboratories, NTT Corporation, Japan

Innovative R&D by NTT

ICASSP 2019

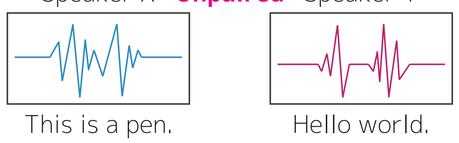
Objective: Non-parallel VC

(Typical) parallel VC: Requires parallel utterances for training



Pros: Easy to learn
Cons: Hard to collect

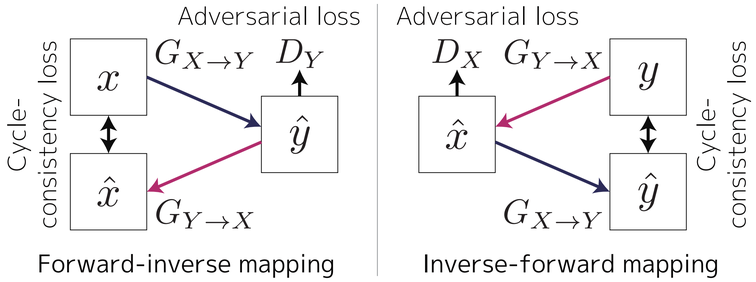
(Our) non-parallel VC: Does not require parallel utterances



Pros: Easy to collect
Cons: Hard to learn
Challenge to address

Baseline: CycleGAN-VC [Kaneko+2017]

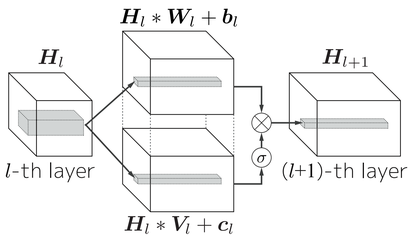
1. CycleGAN [Zhu+2017]



Forward-inverse mapping Inverse-forward mapping

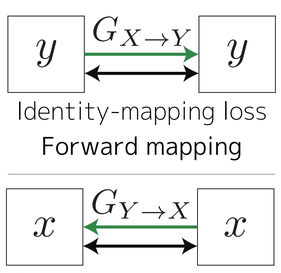
Forward & inverse mappings are simultaneously learned using adversarial loss and cycle-consistency loss
Finds optimal pseudo pair from non-parallel data

2. Gated CNN [Dauphin+2017]



Gated linear unit (GLU)
Propagates information selectively
Represents sequential & hierarchical structures in speech

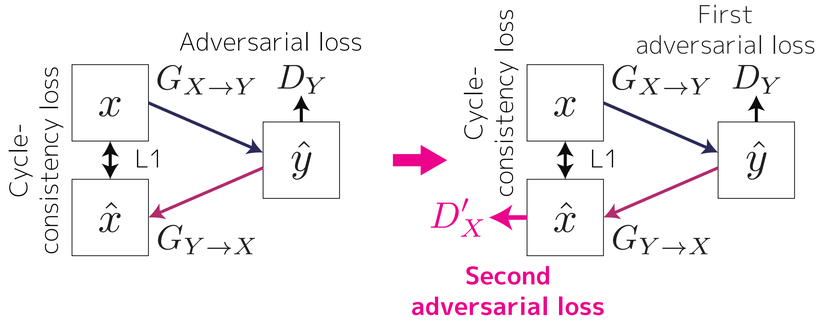
3. IM Loss [Taigman+2017]



Identity-mapping loss
Preserves composition between input and output

Proposed: CycleGAN-VC2

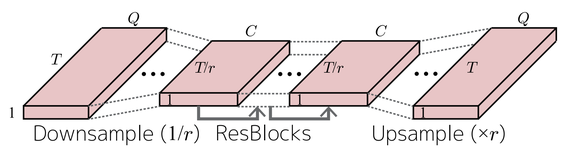
1. Improved objective: Two-step adversarial losses



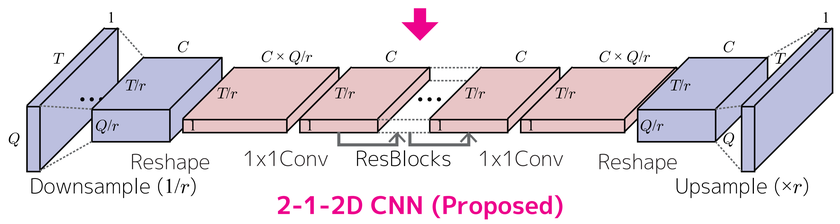
One-step adversarial loss (Previous)
Causes oversmoothing in cyclic reconstruction

Two-step adversarial losses (Proposed)
Mitigates oversmoothing effect over double conversion

2. Improved generator: 2-1-2D CNN



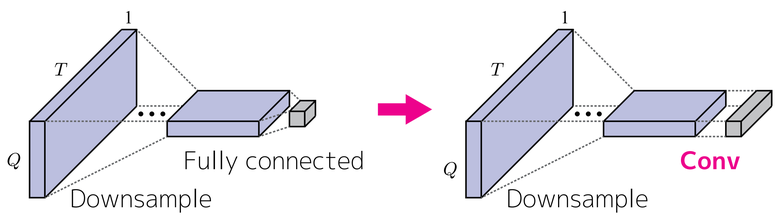
1D CNN (Previous)
Able to translate dynamically but tends to lose original structure



2-1-2D CNN (Proposed)

2D CNN Preserves structure → 1D CNN Translates dynamically → 2D CNN Preserves structure

3. Improved discriminator: PatchGAN [Li+2016]



FullGAN (Previous)
Requires many parameters
Makes training hard

PatchGAN (Proposed)
Reduces parameters
Makes training easy

Experiments

1. Experimental conditions

i) Data

Dataset: Voice Conversion Challenge 2018 (**Spoke** (**non-parallel**) task)
Speakers: Professional US English speakers
Sentences: 81 sentences (about 5 min., **relatively few** for VC)
Sampling Rate: 22.05 kHz
Features: 34 MCEPs, log F_0, APs (WORLD, 5 ms)

ii) Conversion process (Follow VCC 2018 baseline)

Inter-gender: Vocoder-based VC
MCEP: CycleGAN-VC2
log F_0: Linear transformation
AP: No conversion
→ WORLD vocoder [Morise+2016]

Intra-gender: Vocoder-free VC [Kobayashi+2016]
DiffMCEP: CycleGAN-VC2 → Waveform conversion (MLSA filter)

iii) Training

Does not use any extra data, modules, or time alignment procedure

2. Objective evaluation

i) Mel-cepsral distortion (MCD): Global structural difference

Method	SF-TF	SM-TM	SM-TF	SF-TM
CycleGAN-VC2	6.83±.01	6.31±.03	7.22±.05	6.26±.03
CycleGAN-VC†	7.37±.03	6.68±.07	7.68±.05	6.51±.05
Frame-based CycleGAN‡	8.85±.07	7.27±.11	8.86±.27	8.51±.36

ii) Modulation spectra distance (MSD): Local structural difference

Method	SF-TF	SM-TM	SM-TF	SF-TM
CycleGAN-VC2	1.49±.01	1.53±.02	1.45±.00	1.52±.01
CycleGAN-VC†	2.42±.08	2.66±.08	2.21±.13	2.65±.15
Frame-based CycleGAN‡	3.78±.26	2.77±.10	3.32±.06	3.61±.15

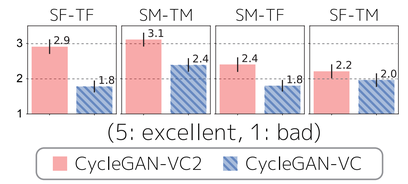
S: Source, T: Target, F: Female, M: Male †[Kaneko+2017] ‡[Fang+2018]

In both metrics, smaller is better

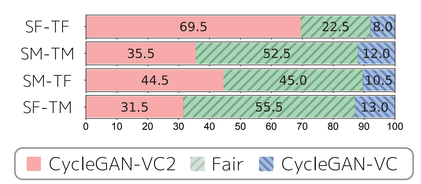
Note: See our paper for detailed ablation studies

3. Subjective evaluation (10 participants)

i) MOS test on naturalness



ii) XAB test on speaker similarity



Improves results for every speaker pair

Listen to speech samples at:

http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc2/index.html

Copyright©2019 NTT corporation All Rights Reserved.