MaskCycleGAN-VC:
Learning Non-parallel Voice Conversion with Filling in Frames
Takuhito Kaneko, Hirokazu Kameoka, Kou Tanaka, Nobukatsu Hojo
NTT Communication Science Laboratories, NTT Corporation, Japan

1 Background and objective
I. Non-parallel voice conversion

- Pros: Easy to collect
- Cons: Hard to learn

(Challenge to be addressed)

II. Non-parallel mel-spectrogram conversion

Recent advances in mel-spectrogram vocoder
- WaveNet [Shen+18], WaveGlow [Prenger+19], MelGAN [Kumar+19], Parallel WaveGAN [Yamamoto+20]

Limited to mel-cepstrum conversion, not mel-spectrogram conversion

Applicable to mel-spectrogram conversion, but requires additional module

As alternative, we propose MaskCycleGAN-VC

III. Challenge of mel-spectrogram conversion

How to convert only voice factors while retaining time-frequency structure in mel-spectrogram?

2 Key idea

Learning non-parallel voice conversion with filling in frames (FIF)

- Strength 1: Additional supervision is not required
- Strength 2: Increase in model size is negligibly small

3 Baseline: CycleGAN-VC2

Learning non-parallel conversion based on cycle consistency

5 Experiments

I. Experimental settings

Dataset: Spoke task of Voice Conversion Challenge 2018 [Lorenzo-Trubar+18]
- Four speakers: VCC2SF3, VCC2SM3, VCC2TF1, and VCC2TM1

Utterances: 81 utterances for training (5 min) & 35 utterances for evaluation

Sampling rate: 22.05 kHz

Conversion target: 80-dimensional log mel-spectrogram

Waveform synthesis: MelGAN vocoder [Kaneko+19]

II. Objective evaluation

Metrics: MCD [dB]/KSD [×10^5] [Binkowski+2020] (Smaller values are preferable)

i. Comparison among different-sized masks

FIF X% (constant) is missing

FIF D-X (variable) is missing

II. Similarity among different types of masks

Subsequent frames

Non-subsequent frames

Point-wise

GLoss

NS

FIF

5. Perform inverse conversion ahead of this process

3. Fill in the missing frames through forward conversion process

4. Generate temporal mask

2. Missing frames

1. m

- Same as CycleGAN-VC2 losses

Cycle-consistency loss

Procedure

1. Converts source mel-spectrogram to target mel-spectrogram

2. Reconstructs source mel-spectrogram from the converted mel-spectrogram

3. Inverse conversion

4. m

Source

Target

CycleGAN-VC2

In CycleGAN-VC2, harmonic structure is compromised

How to prevent?

Preference score [%]

i. AB test on naturalness

ii. XAB test on speaker similarity

Audio samples

http://www.kecl.ntt.co.jp/people/kanekotakuhiro/projects/maskcyclegan-vc/

Copyright 2021 NTT CORPORATION