MaskCycleGAN-VC:
Learning Non-parallel Voice Conversion with Filling in Frames

Audio samples

Takuhiro Kaneko  Hirokazu Kameoka  Kou Tanaka  Nobukatsu Hojo

NTT Communication Science Laboratories, NTT Corporation, Japan
Non-parallel voice conversion

• Training voice converter without parallel corpus

Speaker A
Hello.

Speaker B
Good bye.

Pros: Easy to collect
Cons: Hard to learn (challenge to be addressed)
Non-parallel conversion in mel-spectrogram domain

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

- Recent advances in non-parallel VCs (e.g., CycleGAN-VCs [Kaneko+17/19/20])
  - CycleGAN-VC/VC2: Limited to mel-cepstrum conversion, not mel-spectrogram conversion
  - CycleGAN-VC3: Applicable to mel-spectrogram conversion, but requires additional module

→ As alternative, we propose MaskCycleGAN-VC
Background and Objective 3/3

Challenge of mel-spectrogram conversion

• Required to convert only voice factors while retaining **time-frequency structure**

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Converted by CycleGAN-VC2 (previous)</th>
</tr>
</thead>
</table>

Mel-spectrogram

Frame

Frame

Frame

CycleGAN-VC2 [Kaneko+19] does not have sufficient ability to capture time-frequency structure

→ **Harmonic structure is compromised**

Q. How to prevent?
Key Idea

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

1. Create **missing frames** artificially

2. **Fill in missing frames** based on surrounding frames
→ Learn time-frequency structure in **self-supervised** manner

**Strength 1:** Additional **supervision is not required**

**Strength 2:** **Increase in model size** is negligibly small

Related work
- Representation learning via image inpainting (Context Encoder [Pathak+2016])
- Representation learning via text infilling (MaskGAN [Fedus+2018], BERT [Devlin+2019])
Learning non-parallel conversion based on cycle consistency

- Networks: Converter, inverse converter, discriminator, and second discriminator

The same procedure is used for inverse cycle
Losses: CycleGAN-VC2 is optimized using four losses

1. Cycle-consistency loss
2. Adversarial loss
3. Second adversarial loss
4. Identity-mapping loss

In practice, identity-mapping loss is also used for input preservation.
Proposal: MaskCycleGAN-VC 1/5

Learning non-parallel conversion with filling in frames

Source → Missing frames → Converted → Reconstructed

Mask (w/ missing frames) → Converter → Discriminator → Mask (w/o missing frames) → Inverse converter → Second discriminator
Proposal: MaskCycleGAN-VC 2/5

Learning non-parallel conversion with filling in frames

1. Generate temporal mask

2. Create missing frames artificially

Source → Mask (w/ missing frames) → Missing frames

Reconstructed → Inverse converter → Converted

Discriminator → Second discriminator → Mask (w/o missing frames)
Learning non-parallel conversion with filling in frames

Proposal: MaskCycleGAN-VC 3/5

Fill in through conversion

Source

Missing frames

Converted

Reconstructed

Converter

Discriminator

Inverse converter

Second discriminator

Mask
(w/ missing frames)

Mask
(w/o missing frames)
Proposal: MaskCycleGAN-VC 4/5

Learning non-parallel conversion with filling in frames

Source → Missing frames → Converter → Converted → Reconstructed

④ Assume filling has been accomplished

Mask (w/ missing frames) → Discriminator

⑤ Perform inverse conversion
Proposal: MaskCycleGAN-VC 5/5

Losses: Same as CycleGAN-VC2 losses

1. Cycle-consistency loss
   - Find pseudo pairs
   - Make converted feature appear to be the target
   - Make reconstructed feature appear to be the source

2. Adversarial loss
   - Converter
   - Inverse converter

3. Second adversarial loss
   - Make converted feature appear to be the target
   - Make reconstructed feature appear to be the source

Source → Missing frames → Converted → Reconstructed

Mask (w/ missing frames) → Converter → Inverse converter

Copyright 2021 NTT CORPORATION
Learning non-parallel conversion with filling in frames (FIF)

1. Create **missing frames** artificially
2. Fill in **missing frames** based on surrounding frames
   → Learn time-frequency structure in **self-supervised** manner

**Strength 1:** Additional **supervision is not required**

**Strength 2:** **Increase in model size** is negligibly small

**Related work**
- Representation learning via image inpainting (Context Encoder [Pathak+2016])
- Representation learning via text infilling (MaskGAN [Fedus+2018], BERT [Devlin+2019])
Experimental Settings

Data

• **Dataset:** Spoke task of Voice Conversion Challenge 2018 [Lorenzo-Trueba+18]
  › 4 speakers: VCC2SF3, VCC2SM3, VCC2TF1, & VCC2TM1 (S: Source, T: Target, F: Female, M: Male)
• **Utterances:** 81 utterances for training (5 min) & 35 utterances for evaluation
• **Sampling rate:** 22.05 kHz
• **Conversion target:** 80-dimensional log mel-spectrogram

Conversion and synthesis

Speaker A

Mel-spectrogram extractor

MaskCycleGAN-VC

Converter

Speaker B

MelGAN vocoder [Kumar+19]
Objective Evaluation 1/3

Comparison among different-sized masks

<table>
<thead>
<tr>
<th>Method</th>
<th>SF-TF</th>
<th>SM-TM</th>
<th>SF-TM</th>
<th>SM-TF</th>
</tr>
</thead>
<tbody>
<tr>
<td>① FIF 0</td>
<td>7.66/786</td>
<td>7.11/356</td>
<td>6.91/277</td>
<td>8.11/1094</td>
</tr>
<tr>
<td>② FIF 25</td>
<td>7.45/560</td>
<td>6.85/297</td>
<td>6.76/249</td>
<td>7.84/775</td>
</tr>
<tr>
<td>③ FIF 0-25</td>
<td>7.45/489</td>
<td>6.83/103</td>
<td>6.78/206</td>
<td>7.80/605</td>
</tr>
<tr>
<td>④ FIF 0-50</td>
<td>7.37/467</td>
<td>6.77/83.8</td>
<td>6.73/146</td>
<td>7.64/502</td>
</tr>
<tr>
<td>⑤ FIF 0-75</td>
<td>7.40/468</td>
<td>6.75/89.2</td>
<td>6.72/169</td>
<td>7.66/546</td>
</tr>
</tbody>
</table>

1. Zero-sized (①) vs non-zero sized (②–⑤): Non-zero sized mask is better
2. Constant-sized (②) vs variable-sized (④): Variable-sized mask is better
3. Size dependency (③–⑤): FIF 0-50 is the best

MCD [dB]/KDSD [x10^5]
Smaller values are preferable

Mel-Cepstral Distortion Kernel DeepSpeech Distance
[Binkowski+2020]
**Objective Evaluation 2/3**

**Comparison among different types of masks**

<table>
<thead>
<tr>
<th>Method</th>
<th>SF-TF</th>
<th>SM-TM</th>
<th>SF-TM</th>
<th>SM-TF</th>
</tr>
</thead>
<tbody>
<tr>
<td>① FIF</td>
<td>7.37/467</td>
<td>6.77/83.8</td>
<td>6.73/146</td>
<td>7.64/502</td>
</tr>
<tr>
<td>② FIF&lt;sub&gt;NS&lt;/sub&gt;</td>
<td>7.53/648</td>
<td>7.00/638</td>
<td>6.90/270</td>
<td>7.97/1181</td>
</tr>
<tr>
<td>③ FIS</td>
<td>7.52/727</td>
<td>6.95/437</td>
<td>6.88/418</td>
<td>7.94/974</td>
</tr>
<tr>
<td>④ FIP</td>
<td>7.65/920</td>
<td>6.97/449</td>
<td>7.09/774</td>
<td>8.24/2126</td>
</tr>
</tbody>
</table>

**Mel-Cepstral Distortion** vs **Kernel DeepSpeech Distance**

**MCD [dB]/KDSD [x10^5]**

Smaller values are preferable

- **FIF (①)** is the best
  - **Subsequent temporal mask** is the most useful for helping non-parallel learning
## Objective Evaluation 3/3

### Comparison among CycleGAN-VCs

<table>
<thead>
<tr>
<th>Method</th>
<th>SF-TF</th>
<th>SM-TM</th>
<th>SF-TM</th>
<th>SM-TF</th>
<th>#param</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MaskCycleGAN-VC (proposed)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>① Mask</td>
<td>7.37/467</td>
<td>6.77/83.8</td>
<td>6.73/146</td>
<td>7.64/502</td>
<td>16M</td>
</tr>
<tr>
<td><strong>CycleGAN-VC2 (w/o FIF)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>② V2 [Kaneko+19]</td>
<td>7.66/891</td>
<td>7.07/509</td>
<td>6.96/494</td>
<td>8.07/1107</td>
<td>16M</td>
</tr>
<tr>
<td><strong>CycleGAN-VC3 (latest)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **MaskCycleGAN-VC (①) is the best**
  - In terms of model size, **Mask** is similar to **V2** and smaller than **V3**
Subjective Evaluation

Comparison: ① V2 (w/o FIF) vs Mask (w/ FIF), ② V3 (latest) vs Mask (proposed)

- AB test on naturalness
- XAB test on speaker similarity

• **Mask outperforms V2** & **V3** in terms of both metrics

Audio Samples

Female (SF3) → Male (TM1)

Source  | Target  | V2  | V3  | Mask
---|---|---|---|---

Male (SM3) → Male (TM1)

Source  | Target  | V2  | V3  | Mask
---|---|---|---|---


Audio samples
MaskCycleGAN-VC Search
Summary and Conclusion

Objective
• Non-parallel mel-spectrogram conversion

Proposal
• MaskCycleGAN-VC
  › Learning non-parallel conversion with FIF

Experimental results
• Naturalness & speaker similarity: Mask outperforms V2 & V3
• Model size: Mask is similar to V2 and smaller than V3

Future work
• Applications to multi-domain VC and application-side VC

Audio samples