NTT

Audio samples

MaskCycleGAN-VC Search

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

ICASSP2021 TORONTO Canada June 6-11, 2021 Metro Toronto Convention Centre

MaskCycleGAN-VC:

Learning Non-parallel Voice Conversion with Filling in Frames

Takuhiro Kaneko

Hirokazu Kameoka

Kou Tanaka

Nobukatsu Hojo

NTT Communication Science Laboratories, NTT Corporation, Japan

Background and Objective 1/3

Non-parallel voice conversion

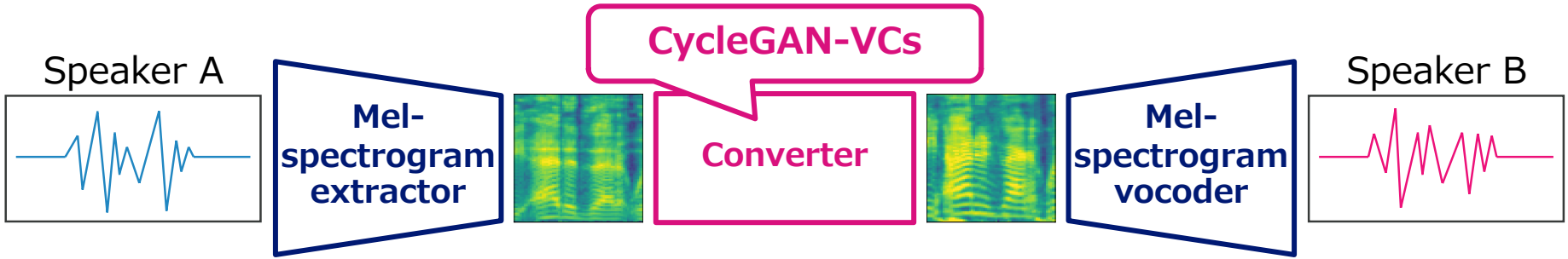
- Training voice converter **without parallel corpus**

Pros: Easy to collect
Cons: Hard to learn **(challenge to be addressed)**

Background and Objective 2/3

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]



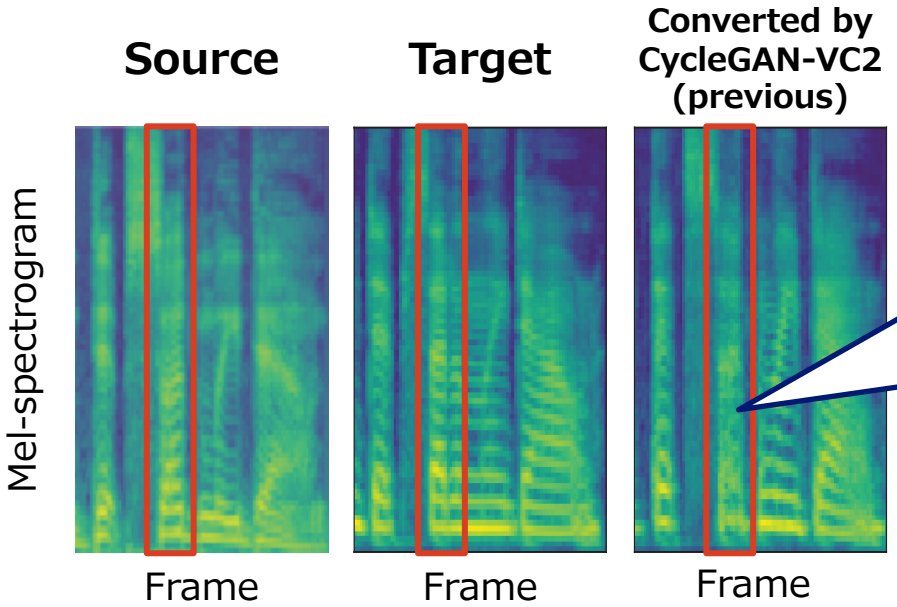
- Recent advances in non-parallel VCs (e.g., **CycleGAN-VCs** [Kaneko+17/19/20])
  - › CycleGAN-VC/VC2: Limited to **mel-cepstrum** conversion, not **mel-spectrogram** conversion
  - › CycleGAN-VC3: Applicable to mel-spectrogram conversion, but requires **additional module**

→ 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**



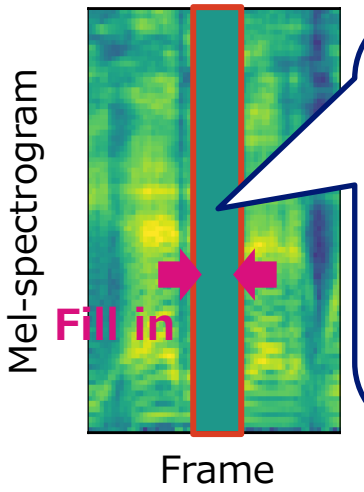
CycleGAN-VC2 [Kaneko+19] does not have sufficient ability to capture time-frequency structure

→ **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**

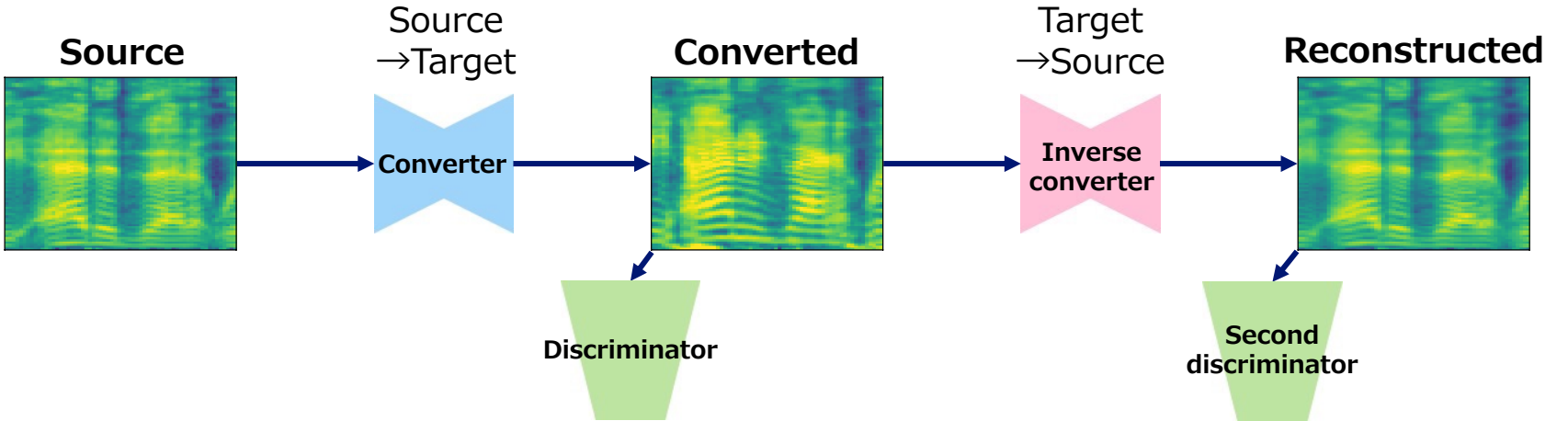
Strength 2: **Increase in model size** is negligibly **small**

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

Baseline: CycleGAN-VC2 [Kaneko+19] 1/2

Learning non-parallel conversion based on cycle consistency

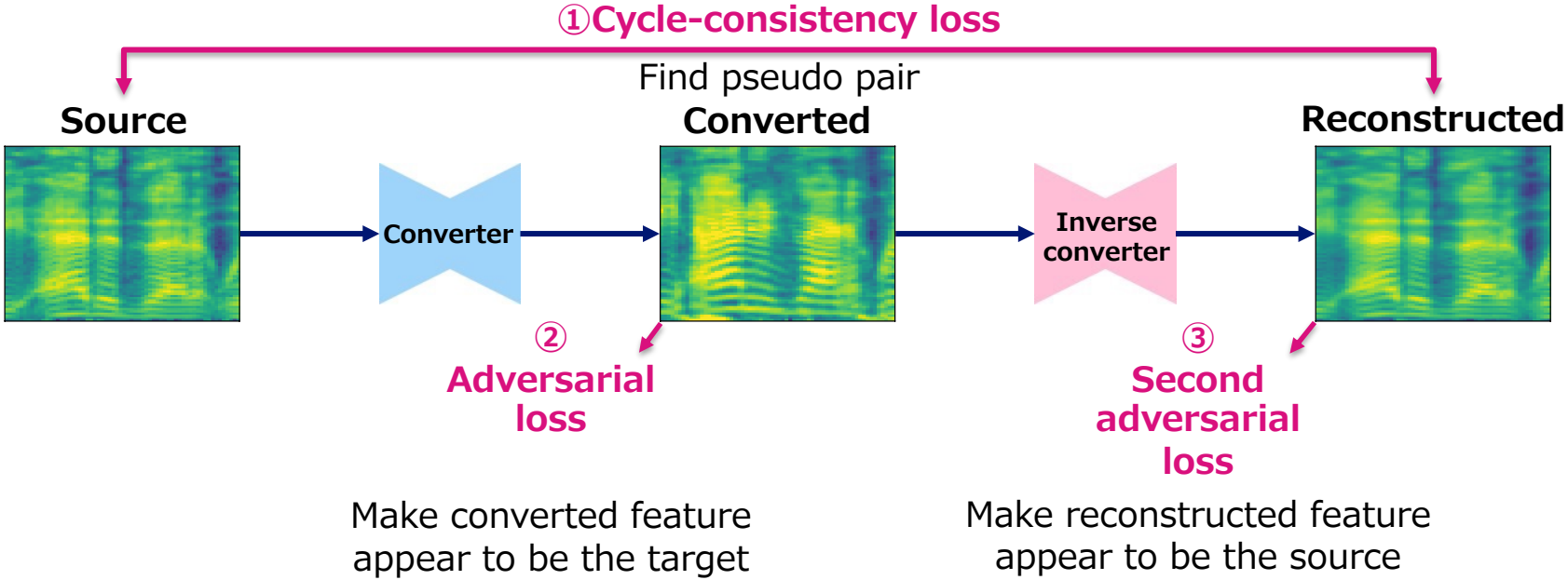
- Networks: Converter, inverse converter, discriminator, and second discriminator



The same procedure is used for inverse cycle

Baseline: CycleGAN-VC2 [Kaneko+19] 2/2

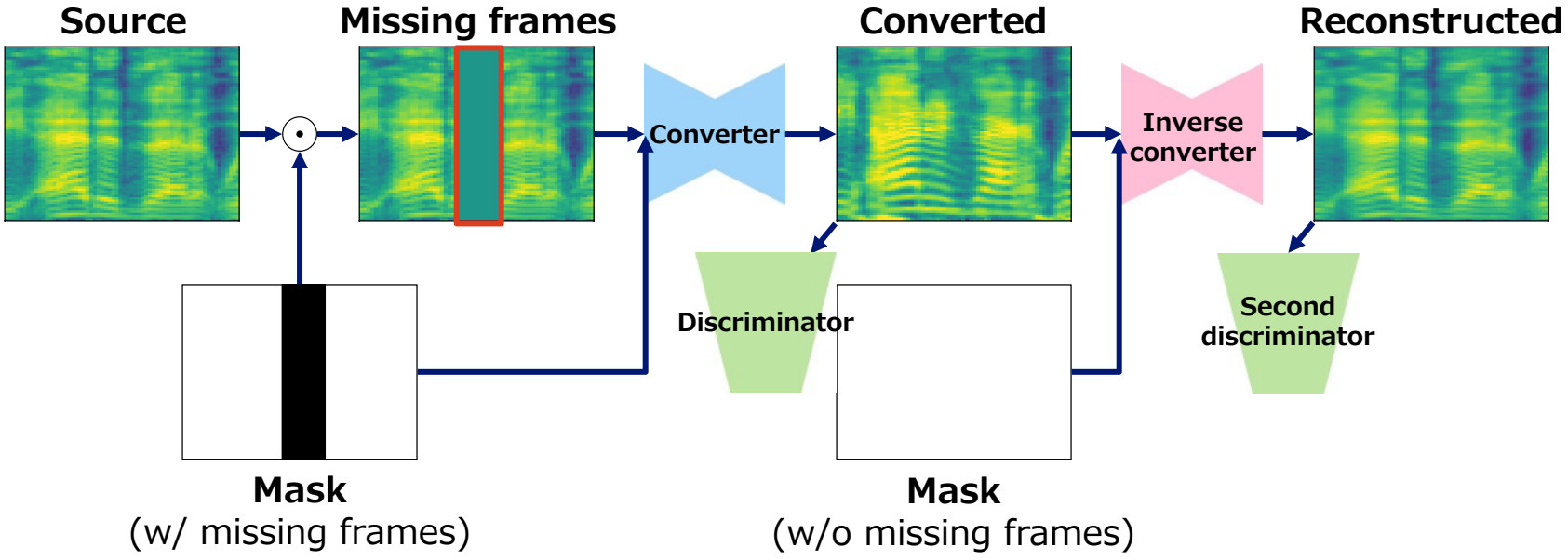
Losses: CycleGAN-VC2 is optimized using four losses



In practice, ④**identity-mapping loss** is also used for input preservation

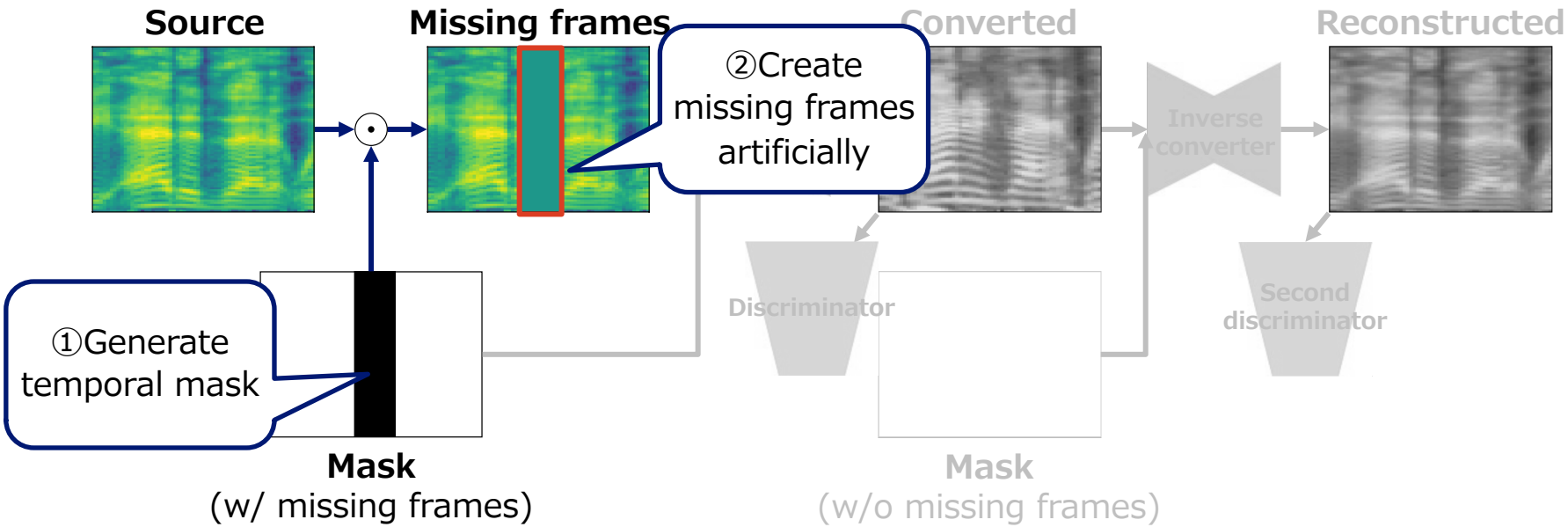
Proposal: MaskCycleGAN-VC 1/5

Learning non-parallel conversion with filling in frames



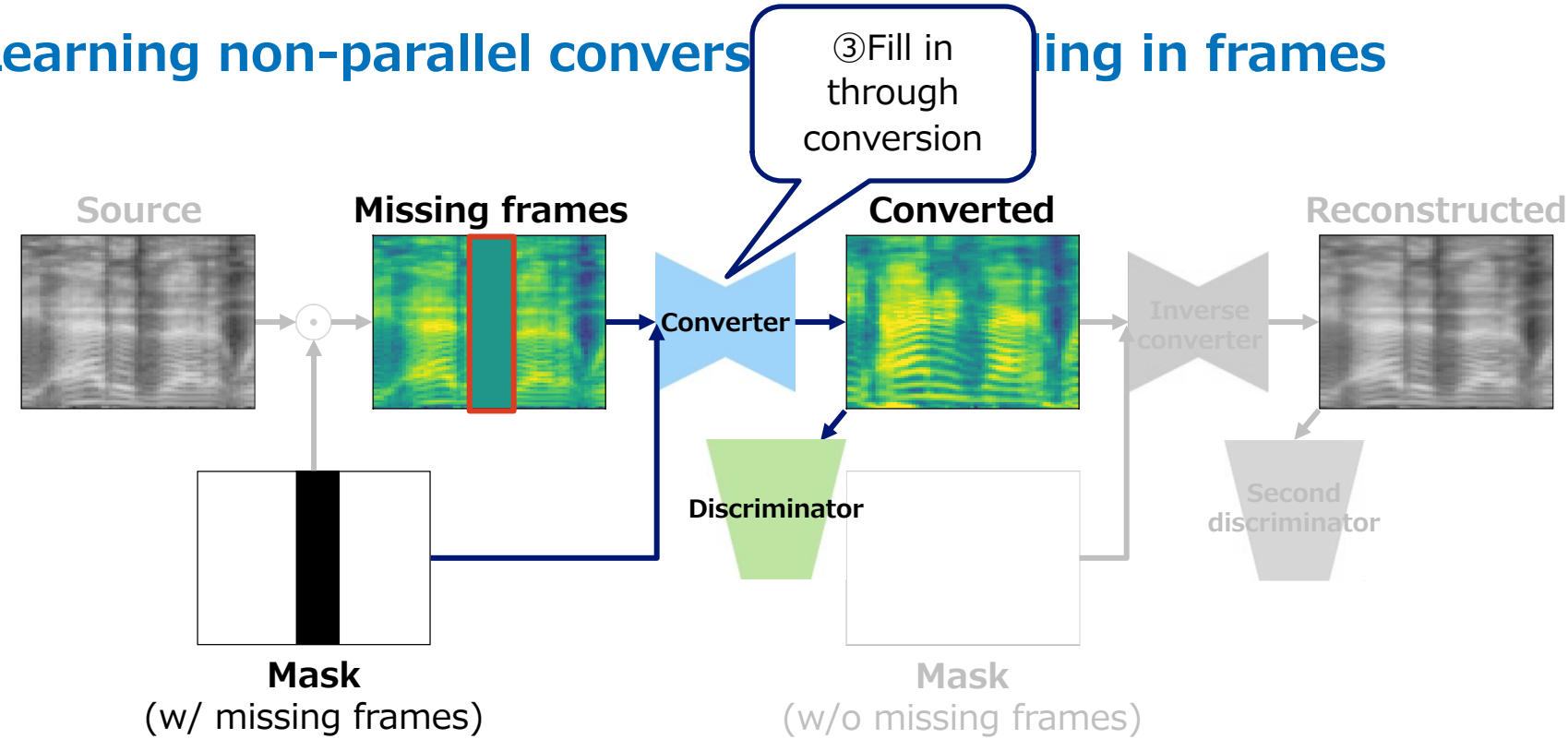
Proposal: MaskCycleGAN-VC 2/5

Learning non-parallel conversion with filling in frames



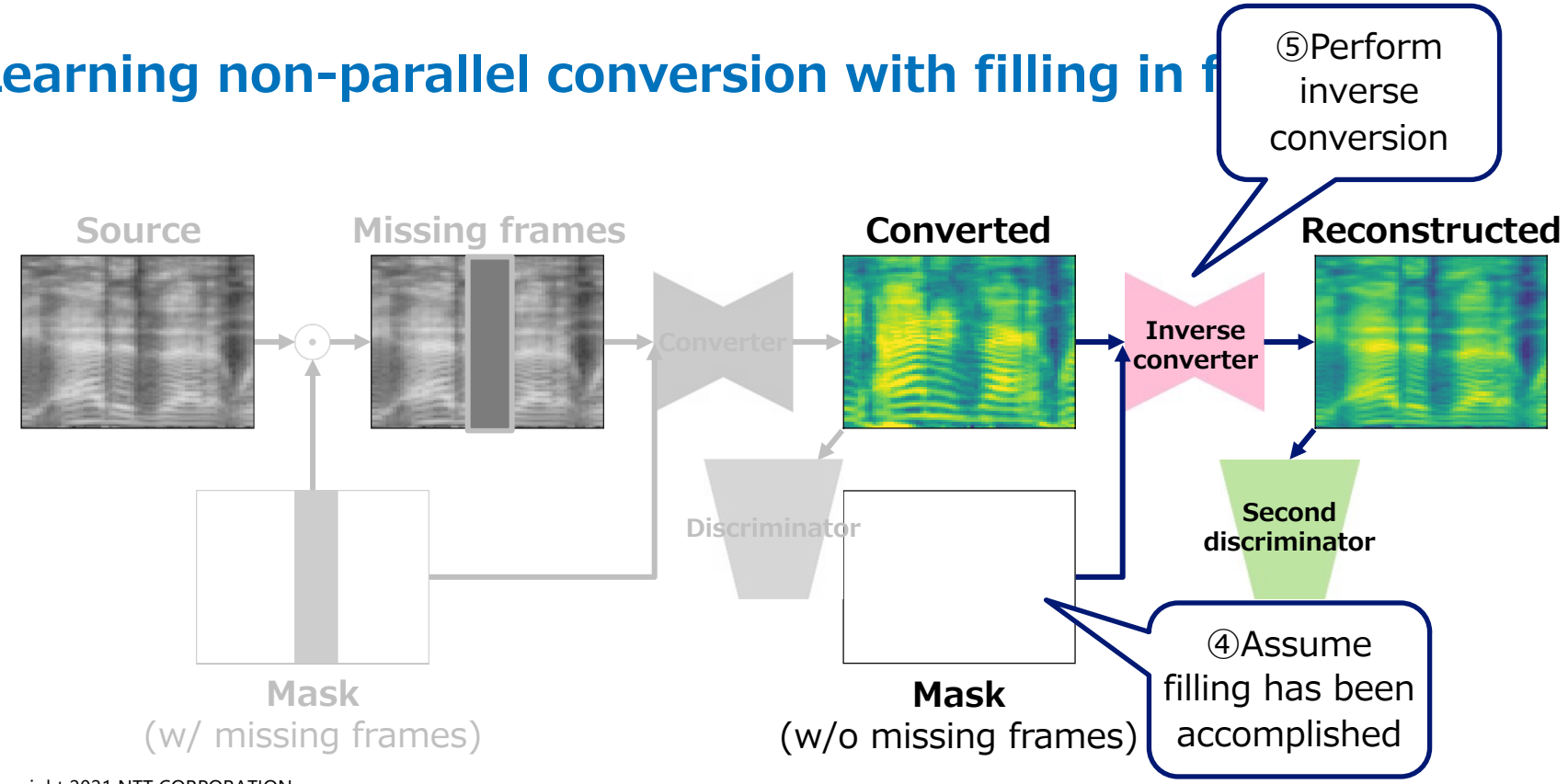
Proposal: MaskCycleGAN-VC 3/5

Learning non-parallel convers ing in frames



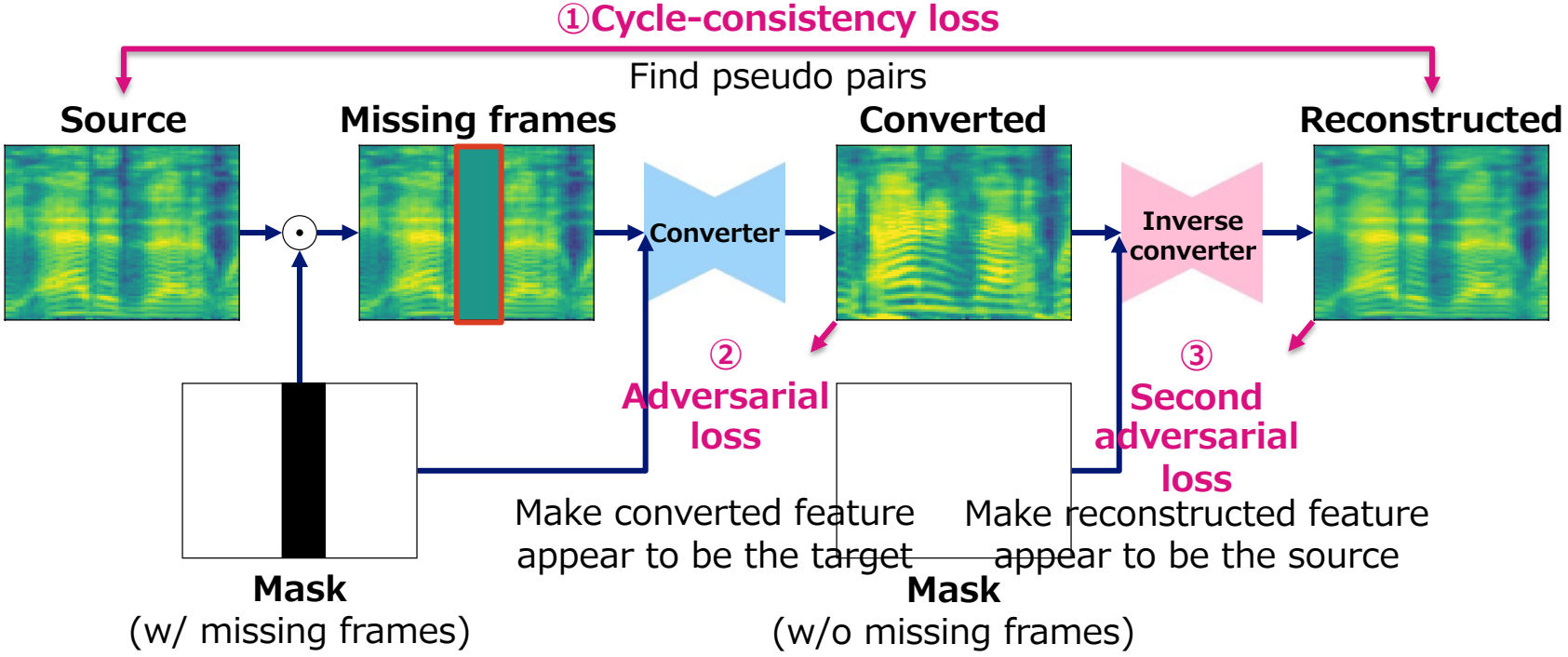
Proposal: MaskCycleGAN-VC 4/5

Learning non-parallel conversion with filling in f



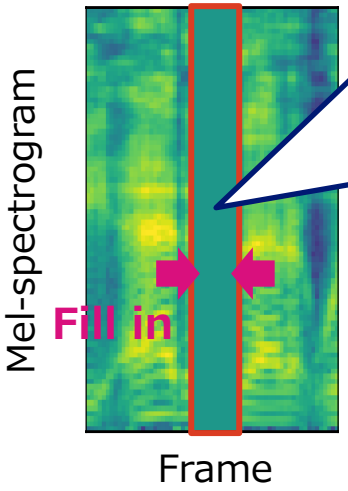
Proposal: MaskCycleGAN-VC 5/5

Losses: Same as CycleGAN-VC2 losses



Key Idea (Reprint)

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**

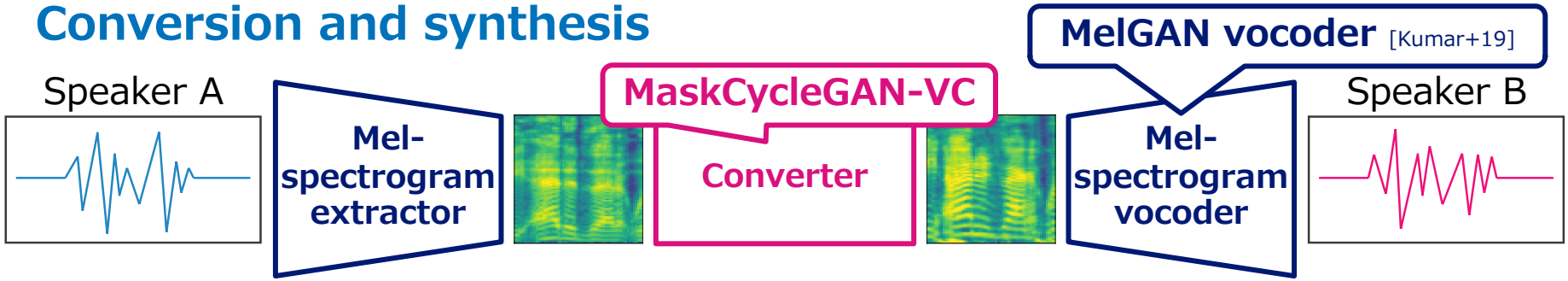
Related work
- Representation learning via image inpainting (Context Encoder [Pathak+2016])
- Representation learning via text infilling (MaskGAN [Fedus+2018], BERT [Devlin+2019])

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



Objective Evaluation 1/3

Mel-**C**epstral **D**istortion **K**ernel **D**eep**S**peech **D**istance [Binkowski+2020]

Comparison among different-sized masks

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

	Method	SF-TF	SM-TM	SF-TM	SM-TF
FIF X: X% (constant) is missing	①FIF 0	7.66/786	7.11/356	6.91/277	8.11/1094
	②FIF 25	7.45/560	6.85/297	6.76/249	7.84/775
FIF 0-X: 0-X% (variable) is missing	③FIF 0-25	7.45/489	6.83/103	6.78/206	7.80/605
	④FIF 0-50	**7.37**/**467**	6.77/**83.8**	6.73/**146**	**7.64**/**502**
	⑤FIF 0-75	7.40/468	**6.75**/89.2	**6.72**/169	7.66/546

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**

Objective Evaluation 2/3

Mel-**C**epstral **D**istortion **K**ernel **D**eep**S**peech **D**istance [Binkowski+2020]

Comparison among different types of masks

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

	Method	SF-TF	SM-TM	SF-TM	SM-TF
Subsequent frames	**①FIF**	**7.37/467**	**6.77/83.8**	**6.73/146**	**7.64/502**
Non-subsequent frames	②FIF_{NS}	7.53/648	7.00/638	6.90/270	7.97/1181
Subsequent spectrogram	③FIS	7.52/727	6.95/437	6.88/418	7.94/974
Point-wise	④FIP	7.65/920	6.97/449	7.09/774	8.24/2126

- **FIF (①) is the best**
  - › **Subsequent temporal mask** is the most useful for helping non-parallel learning

Objective Evaluation 3/3

Mel-**C**epstral **D**istortion **K**ernel **D**eep**S**peech **D**istance [Binkowski+2020]

Comparison among CycleGAN-VCs

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

	Method	SF-TF	SM-TM	SF-TM	SM-TF	#param
MaskCycleGAN-VC (proposed)	**①Mask**	**7.37**/467	**6.77/83.8**	**6.73/146**	**7.64/502**	**16M**
CycleGAN-VC2 (w/o FIF)	②V2 [Kaneko+19]	7.66/891	7.07/509	6.96/494	8.07/1107	**16M**
CycleGAN-VC3 (latest)	③V3 [Kaneko+20]	7.54/**369**	7.10/227	6.91/311	7.97/819	27M

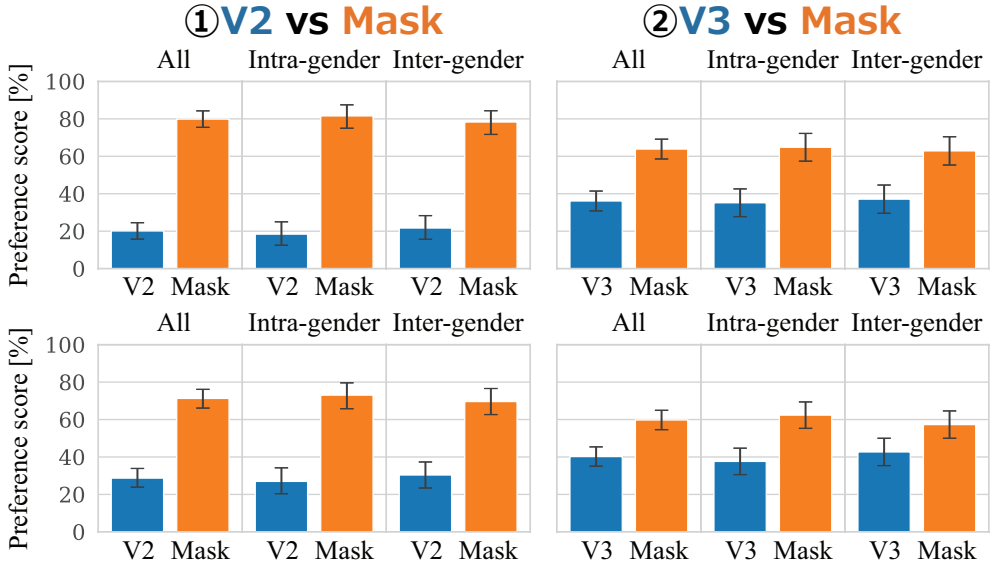
- **MaskCycleGAN-VC (①) is the best**
  - › In terms of model size, **Mask** is similar to **V2** and smaller than **V3**

Subjective Evaluation

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

AB test on naturalness

XAB test on speaker similarity



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

V2: CycleGAN-VC2 [Kaneko+19] V3: CycleGAN-VC3 [Kaneko+20] Mask: MaskCycleGAN-VC (Proposed)

Audio Samples

Audio samples

MaskCycleGAN-VC Search

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

NTT

Female (SF3) → Male (TM1)

Source	Target	V2	V3	Mask
				

Male (SM3) → Male (TM1)

Source	Target	V2	V3	Mask
				

V2: CycleGAN-VC2 [Kaneko+19] **V3**: CycleGAN-VC3 [Kaneko+20] **Mask**: MaskCycleGAN-VC (Proposed)

Summary and Conclusion

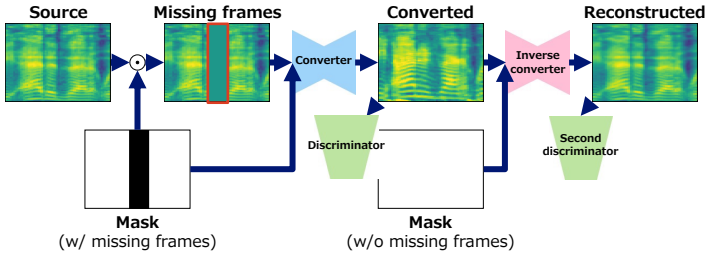
Objective

- Non-parallel mel-spectrogram conversion

Proposal

- **MaskCycleGAN-VC**
  - › Learning non-parallel conversion with **FIF**

MaskCycleGAN-VC



Experimental results

- Naturalness & speaker similarity: **Mask** outperforms **V2** & **V3**
- Model size: **Mask** is similar to **V2** and smaller than **V3**

Audio samples

MaskCycleGAN-VC Search

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

Future work

- Applications to multi-domain VC and application-side VC