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

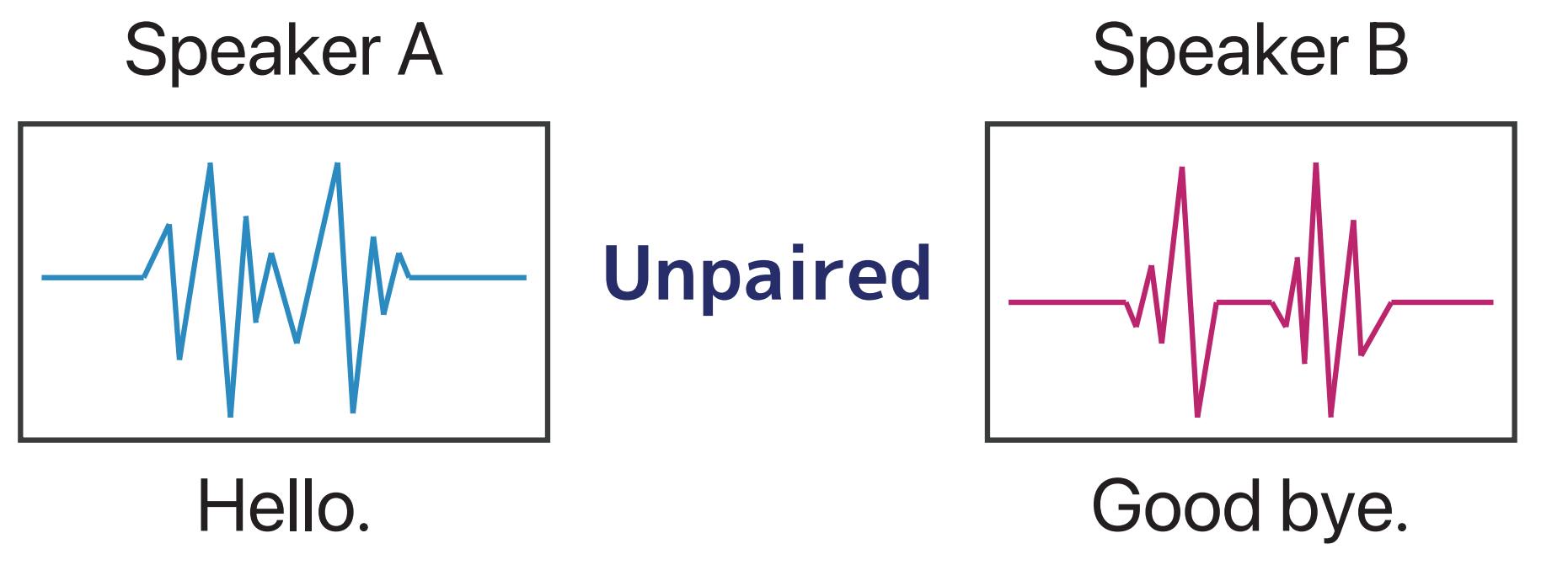
SPE-11: Voice Conversion 1: Non-parallel Conversion

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

1 Background and objective

I. Non-parallel voice conversion

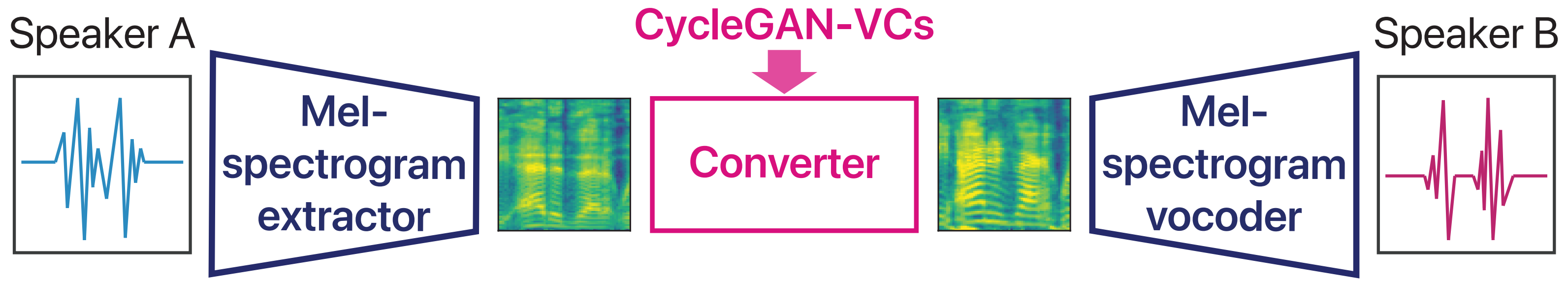


Pros: Easy to collect
Cons: Hard to learn
(Challenge to be addressed)

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]



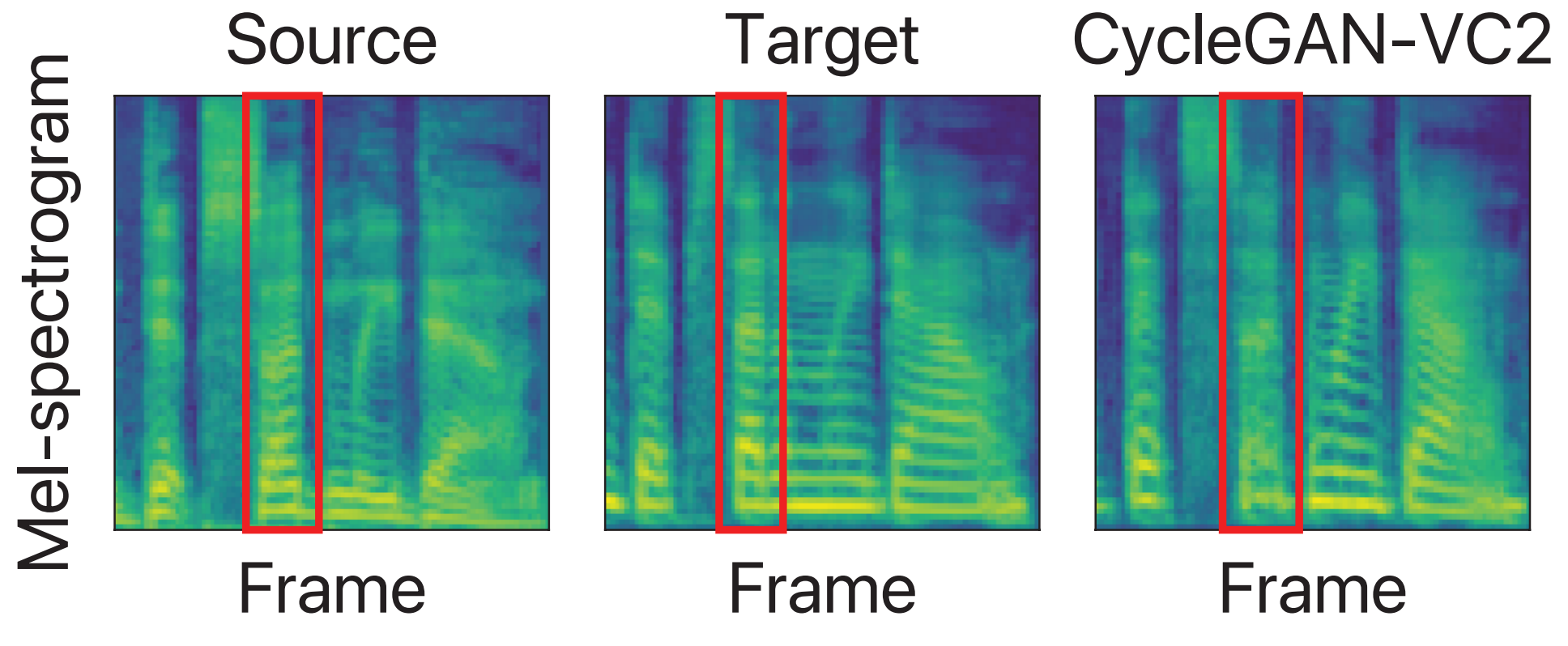
Recent advances in non-parallel VCs (e.g., CycleGAN-VCs)

- **CycleGAN-VC/VC2** [Kaneko+17/19]
  Limited to **mel-cepstrum conversion**, not **mel-spectrogram conversion**
- **CycleGAN-VC3** [Kaneko+20]
  Applicable to mel-specgrogram conversion, but requires **additional module**

→ As alternative, we propose **MaskCycleGAN-VC3**

III. Challenge of mel-spectrogram conversion

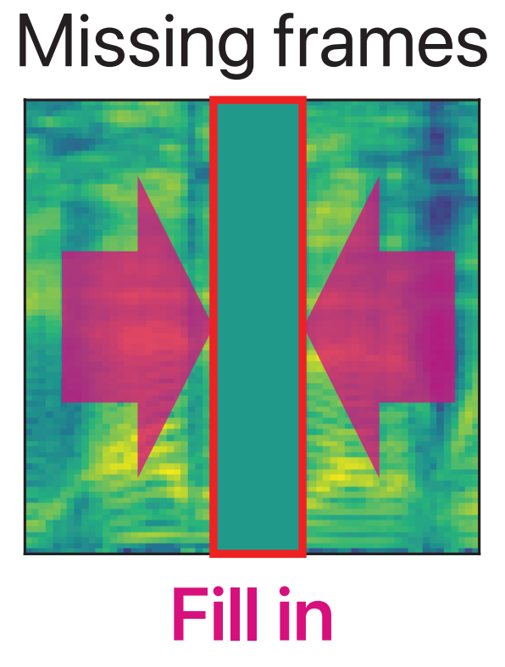
How to convert only voice factors while retaining time-frequency structure in mel-spectrogram?



In CycleGAN-VC2, harmonic structure is compromised
→ **How to prevent?**

2 Key idea

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

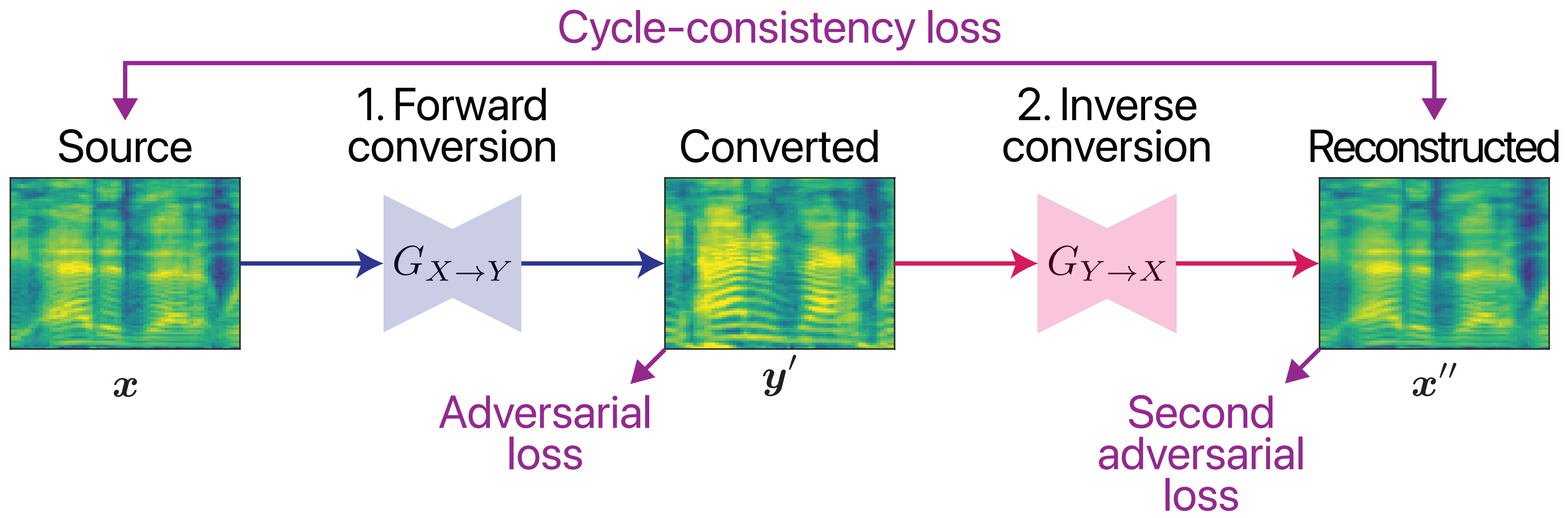


1. Create **missing frames** artificially
2. **Fill in missing frames** based on surrounding frames
→ Learn time-frequency structure in **self-supervised** manner

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

3 Baseline: CycleGAN-VC2 [Kaneko+19]

Learning non-parallel conversion based on cycle consistency



Procedure

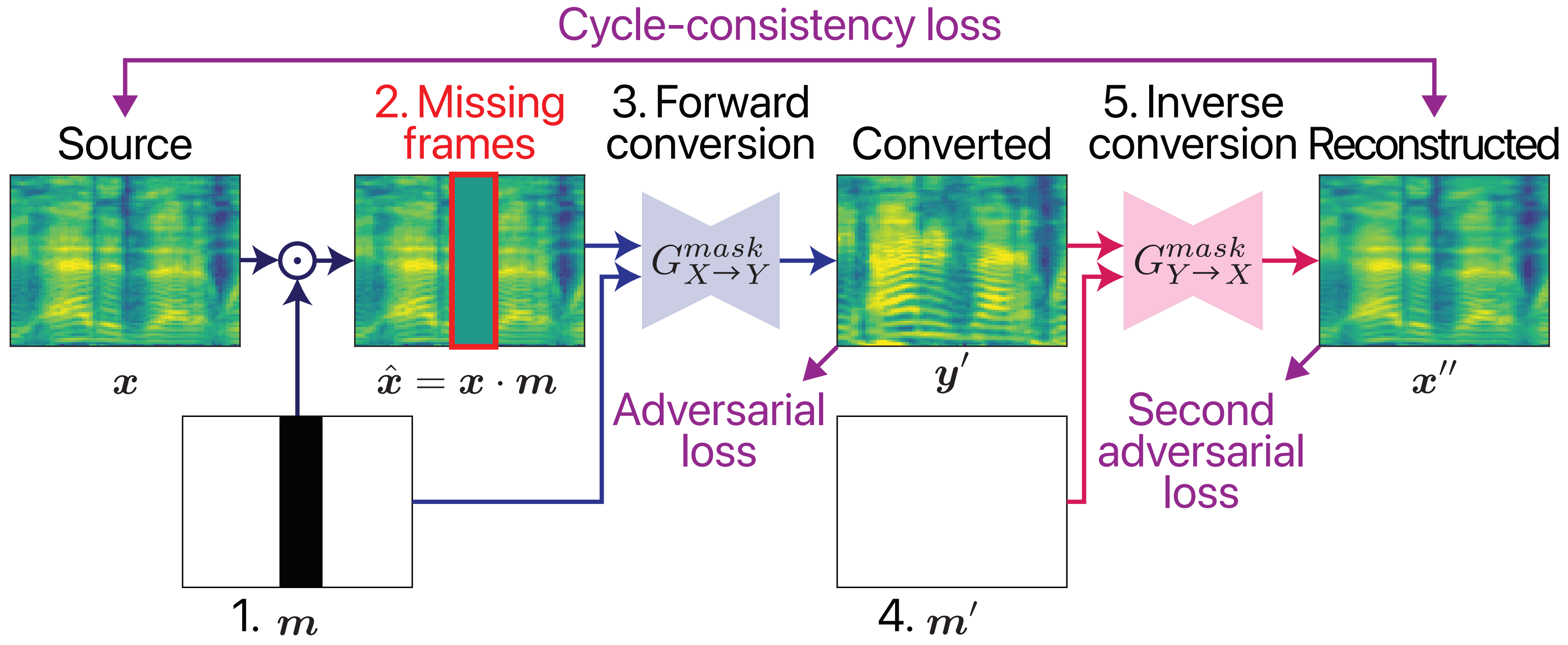
1. Converts source mel-spectrogram to target mel-spectrogram
2. Reconstructs source mel-spectrogram from the converted mel-spectrogram

Losses

- Cycle-consistency loss
  Helps find pseudo pair within cycle-consistency constraint
- Adversarial loss
  Makes converted mel-spectrogram appear to be the target
- Second adversarial loss
  Makes reconstructed mel-spectrogram appear to be the source

4 Proposal: MaskCycleGAN-VC

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



Procedure

1. Generate temporal mask
2. Apply the mask to source mel-spectrogram
   → Create missing frames artificially
3. Fill in the missing frames through forward conversion process
4. Prepare all-ones mask under assumption that filling has been accomplished ahead of this process
5. Perform inverse conversion

Losses

- Same as CycleGAN-VC2 losses

5 Experiments

I. Experimental settings

Dataset: Spoke task of Voice Conversion Challenge 2018 [Lorenzo-Trueba+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 [Kumar+19]

II. Objective evaluation

Metrics: MCD [dB]/**KDSD** [$\times 10^5$] [Binkowski+2020] (Smaller values are preferable)

i. Comparison among different-sized mask

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

ii. Comparison among different types of masks

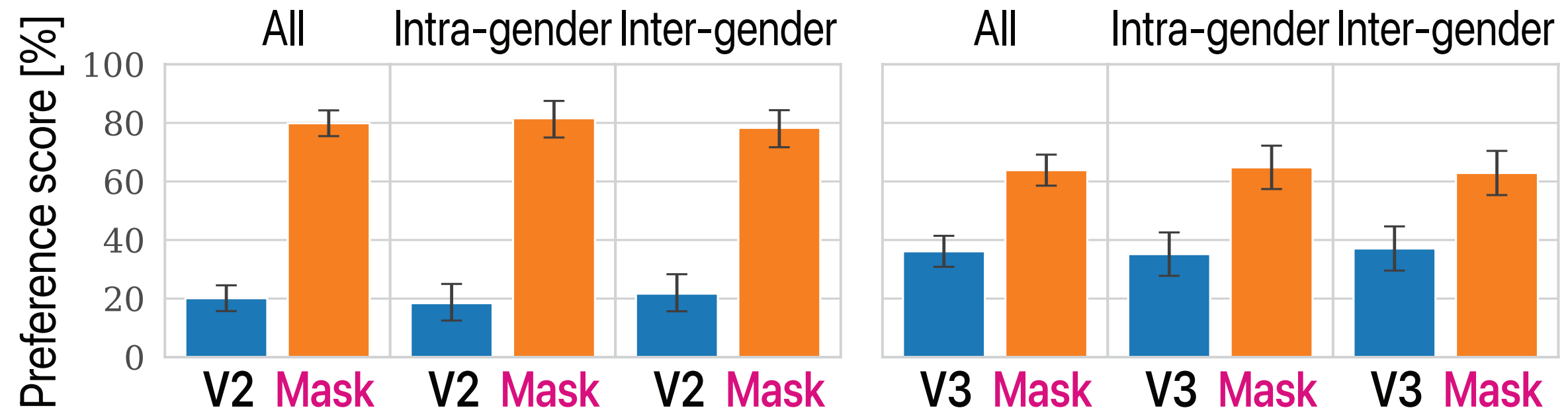
	Type	SF-TF	SM-TM	SF-TM	SM-TF
Subsequent frames	FIF	7.37/467	6.77/83.8	6.73/146	7.64/502
Non-subsequent frames	FIF_{NS}	7.53/648	7.00/638	6.90/270	7.97/1181
Subsequent spectrogram	FIS	7.52/727	6.95/437	6.88/418	7.94/974
Point-wise	FIP	7.65/920	6.97/449	7.09/774	8.24/2126

iii. Comparison among CycleGAN-VCs

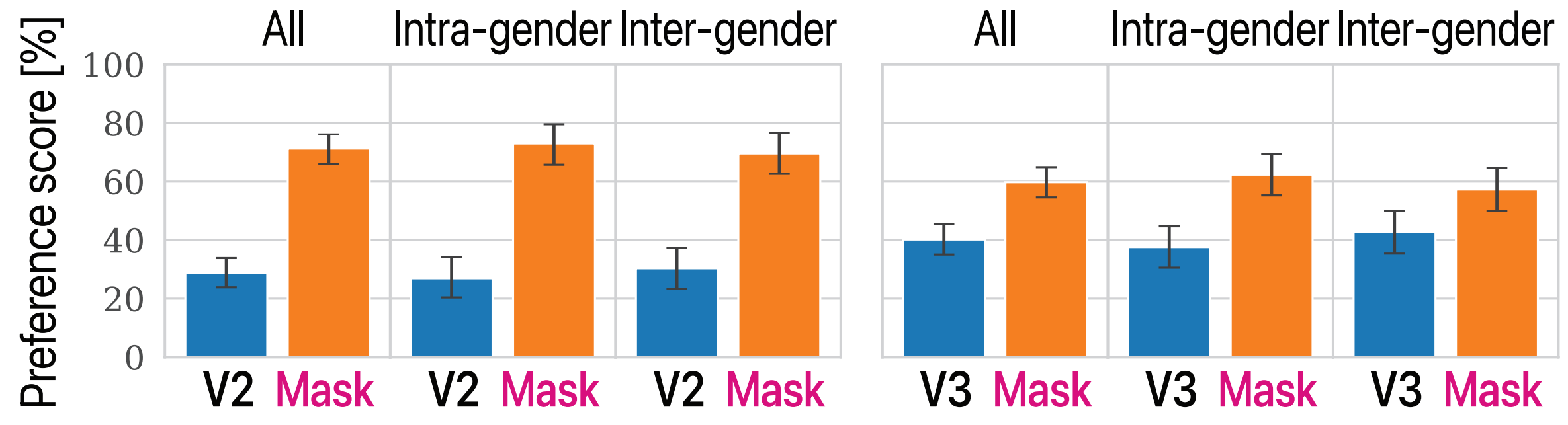
	Model	SF-TF	SM-TM	SF-TM	SM-TF	#param
MaskCycleGAN-VC (Proposed)	Mask	7.37/467	6.77/83.8	6.73/146	7.64/502	16M
CycleGAN-VC2 [Kaneko+19]	V2	7.66/891	7.07/509	6.96/494	8.07/1107	16M
CycleGAN-VC3 [Kaneko+20]	V3	7.54/369	7.10/227	6.91/311	7.97/819	27M

III. Subjective evaluation

i. AB test on naturalness



ii. XAB test on speaker similarity



6 Audio samples

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