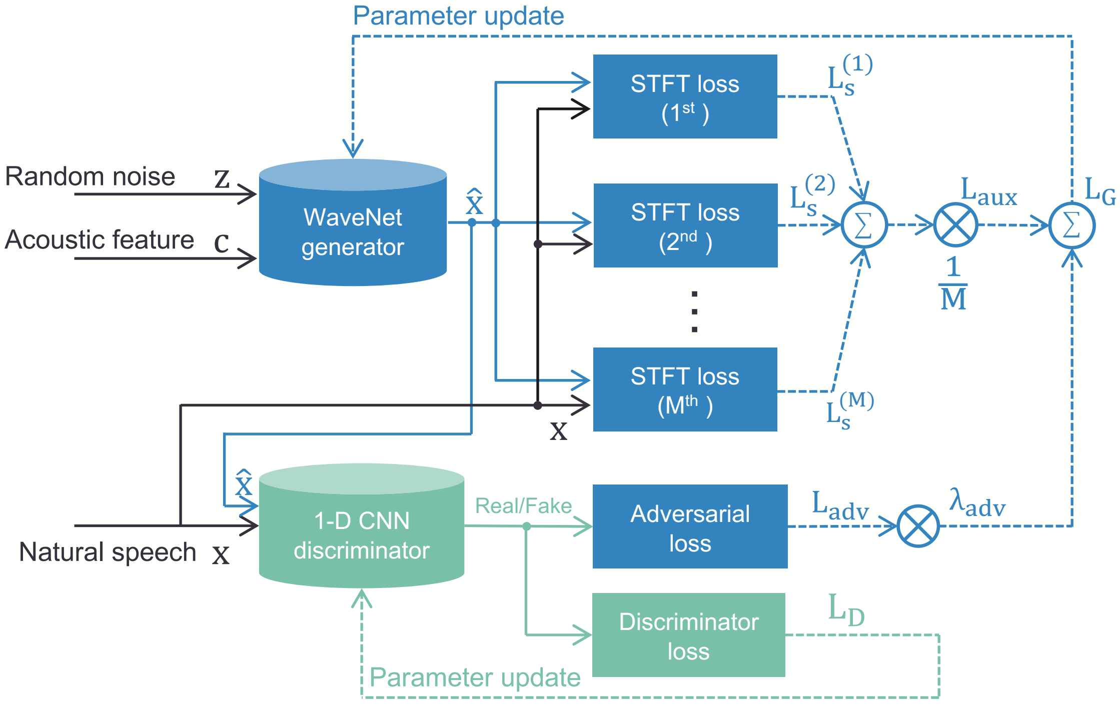
ICASSP2021
TORONTO
Canada
June 6-11, 2021
Metro Toronto Convention Centre

Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators

Ryuichi Yamamoto[1], *Eunwoo Song*[2], *Min-Jae Hwang*[3], *Jae-min Kim*[2]
[1]LINE Corp., Tokyo, Japan
[2]NAVER Corp., Seongnam, Korea
[3]Search Solutions Inc., Seongnam, Korea

LINE
NAVER
NAVER
CLOVA

Parallel WaveGAN (PWG)



Distillation-free

Distillation-free training combining multi-resolution STFT loss and adversarial loss.

High-quality

Competitive perceptual quality to the conventional Parallel WaveNet

Fast

Training and inference speed is much faster than Parallel WaveNet.

Limitation

A single discriminator may not be sufficient to discriminate complex nature of speech.

R. Yamamoto, *et al.*, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.

Overview of our research

Problem

Insufficient capability of the conventional PWG's *discriminator*

Proposed method

Voicing-aware discriminator: separate discriminators for voiced and unvoiced segments

Results

Significant performance improvements for speaker-independent modeling.

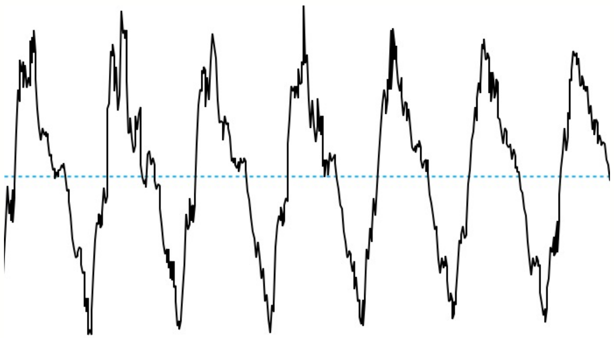
Model	MOS
PWG	3.70 ± 0.05
PWG-V/UV-D (proposed)	$\mathbf{4.23 \pm 0.05}$
Recordings	4.64 ± 0.04

NOTE: MOS in the table was averaged among four speakers. See per-speaker MOS in our paper.

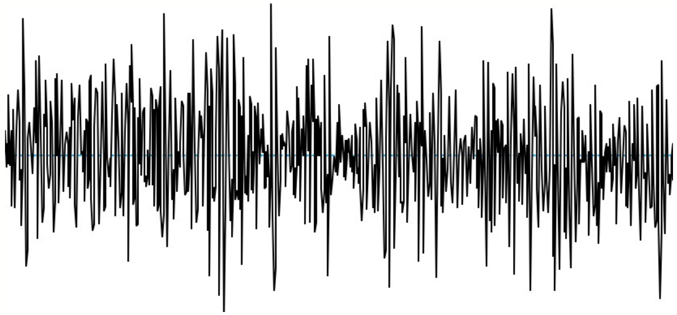
R. Yamamoto, *et al*., "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.

Voiced / unvoiced sounds

/e/



/sh/



Voiced sounds

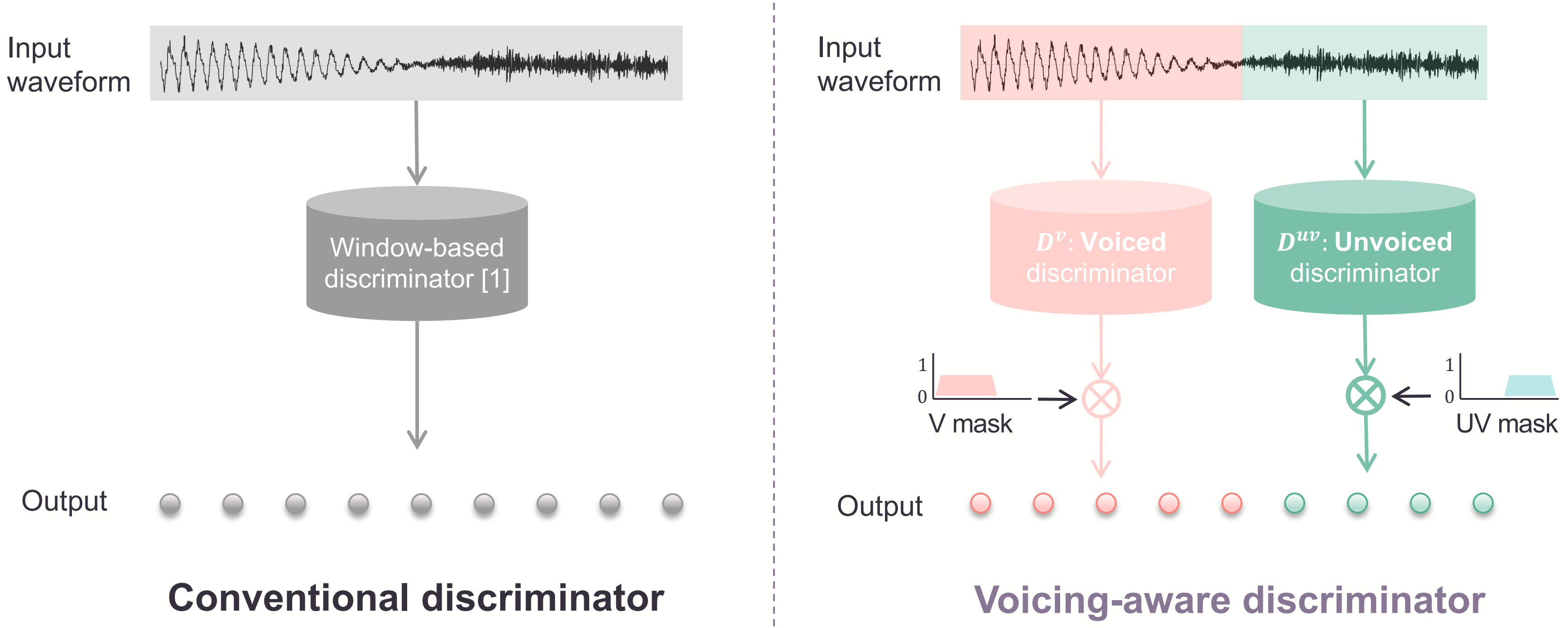
Quasi-periodic (mostly characterized by fundamental frequency and its harmonics)

Unvoiced sounds

Non-periodic (contains noise)

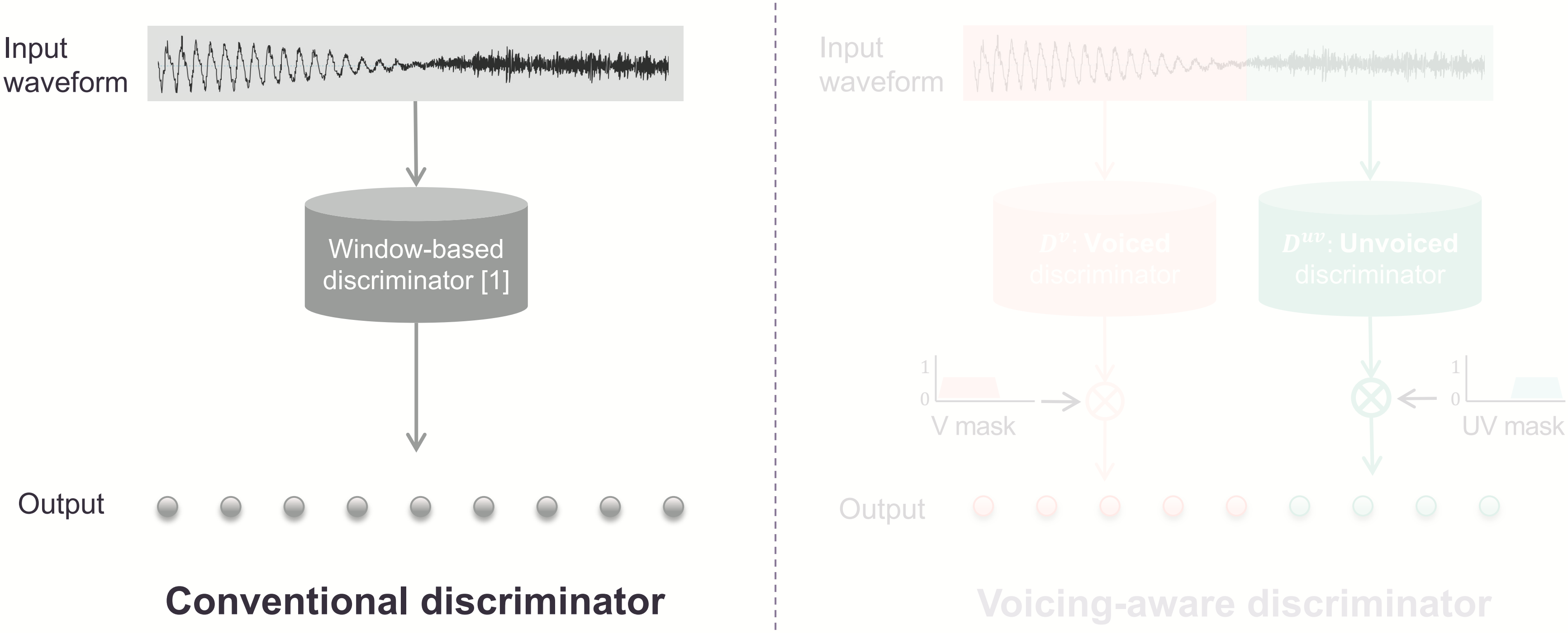
R. Yamamoto, *et al.*, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.

Overview of PWG's discriminators



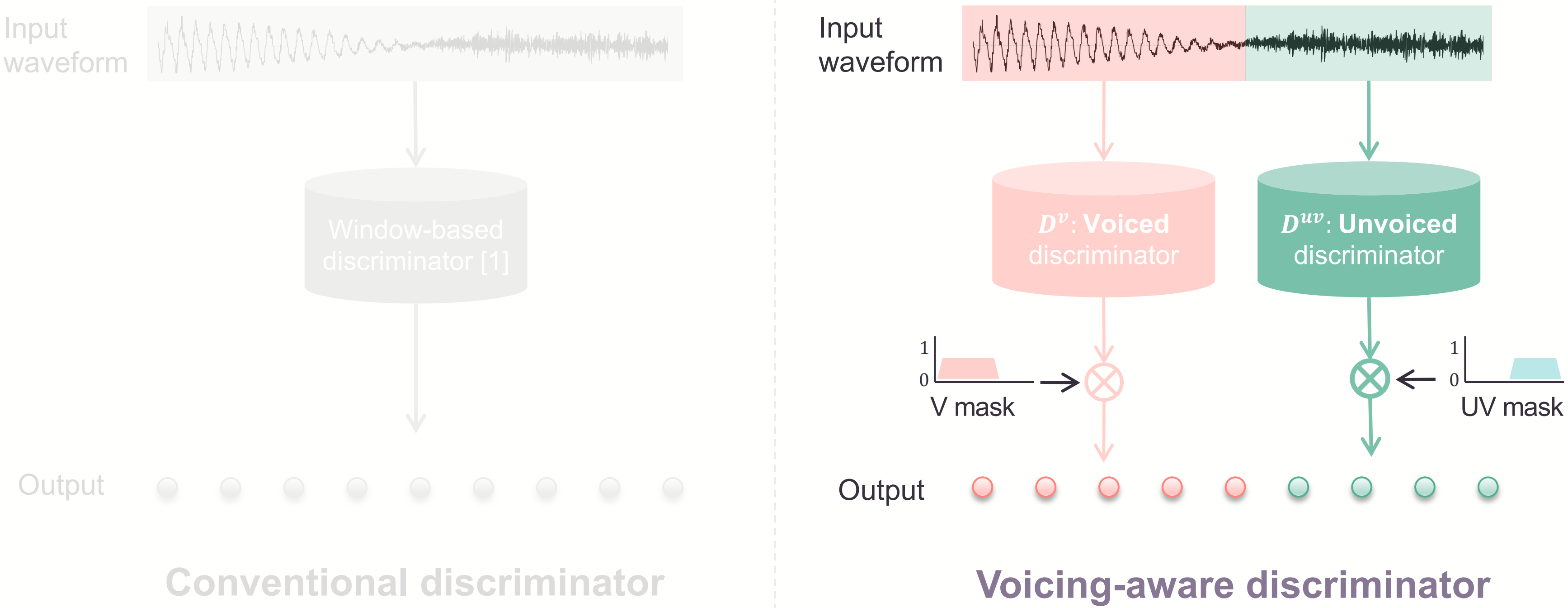
[1] P. Isola, *et al.*, "Image-to-image translation with conditional adversarial networks." in *Proc. CVPR*, 2017.
R. Yamamoto, *et al.*, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.

Overview of PWG’s discriminators



[1] P. Isola, *et al.*, “Image-to-image translation with conditional adversarial networks.” in *Proc. CVPR*, 2017.
R. Yamamoto, *et al.*, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.

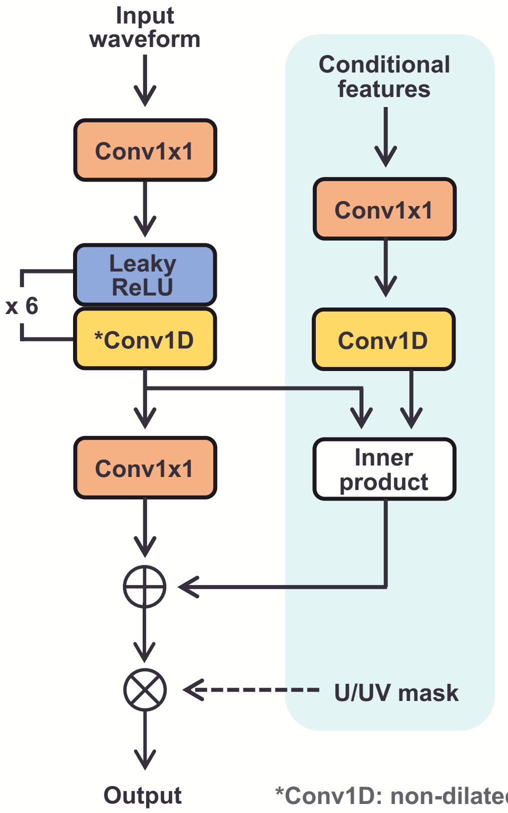
Overview of PWG's discriminators



[1] P. Isola, *et al.*, "Image-to-image translation with conditional adversarial networks." in *Proc. CVPR*, 2017.

R. Yamamoto, *et al.*, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.

Details of voicing-aware discriminator



*Conv1D: non-dilated conv. for D^v

Architecture

- 1-D CNNs
- Conditional discriminator [2]

Designs for voiced/unvoiced discriminators

- Voiced: dilated convolution to increase receptive field
- Unvoiced: non-dilated convolution

Discriminator	Dilation factors	Receptive field
D^{v}	[1, 2, 4, 8, 16, 32]	127
D^{uv}	[1, 1, 1, 1, 1, 1]	13

[2] T. Miyato, *et al.*, “cGANs with projection discriminator,” in *Proc. ICLR*, 2018.
R. Yamamoto, *et al.*, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.

Training objectives

- Least squares GAN (LSGAN [3]) formulation
- Multi-resolution STFT loss (L_{mr_stft}) is used

$$\min_{D} \mathbb{E}_{x,h}\left[\left[(1 - D(x,h))^2\right] + \mathbb{E}_{z,h}\left[D(G(z,h),h)^2\right], \forall D \in \{D^v, D^{uv}\}\right.$$

$$\min_{G} \mathbb{E}_{x,z,h}\left[L_{mr\_stft}(x, G(z,h)\right] + \frac{1}{2}\lambda_{adv}\mathbb{E}_{z,h}\left[\sum_{D\in\{D^v,D^{uv}\}} \left(1 - D(G(z,h),h)\right)^2\right]$$

Auxiliary loss:
Learn from data

Adversarial loss:
Learn from voicing-aware discriminators

x, z, h: waveform, noise, and acoustic features

[3] M. Xudong, *et al.* "Least squares generative adversarial networks." in *Proc. ICCV*, 2017.
R. Yamamoto, *et al.*, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.

Experiments

1. Experiments on discriminator design choices in analysis-by-synthesis
2. Text-to-speech (TTS)
   - FastSpeech 2 [4] is used as an acoustic model.

[4] Ren *et al*., "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *Proc. ICLR,* 2021.
R. Yamamoto, *et al*., "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.

Experimental setup

Data & features

Recordings	Size (training / validation / test)
Two male (M1, M2) and two female (F1, F2) Japanese speakers 24 kHz /16 bit	4,500 (about. 5.5 hours), 250, 250 (per speaker)

Auxiliary features	Frame shift
79-dim ITFTE vocoder parameters [5] (LSFs, log F0, energy, V/UV, REW, SEW)	5 ms

Baseline vocoders

WaveNet [6]

PWG with different discriminator setups (*NOTE*: generator configurations were all the same)

Listening tests

Mean-option score (MOS) listening test on quality and naturalness

Seventeen native Japanese speakers / 20 random utterances for each method

[5] E. Song, *et al*., "Effective spectral and excitation modeling techniques for LSTM-RNN based speech synthesis systems," *IEEE/ACM TASLP*, 2017.
[6] W. Ping, *et al.*, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICLR*, 2019.

MOS test results on analysis-by-synthesis

System	Model	Voiced segments	Unvoiced segments	Discriminator conditioning	MOS
S1	WaveNet	-	-	-	3.48±0.06
S2	PWG	-	-	-	3.59±0.06
S3	PWG-cGAN-D	-	-	Yes	3.97±0.05
S4	PWG-V/UV-D	D^{v}	D^{v}	Yes	3.50±0.06
S5	PWG-V/UV-D	D^{uv}	D^{v}	Yes	3.46±0.05
S6	PWG-V/UV-D	D^{uv}	D^{uv}	Yes	3.64±0.05
S7	**PWG-V/UV-D (proposed)**	D^{v}	D^{uv}	Yes	**4.07±0.05**
R1	Recordings	-	-	-	4.64±0.04

+ 0.38

+ 0.48

NOTE: MOS in the table was averaged among four speakers. See per-speaker MOS in our paper.

: Misconfigured

S2 vs. S3

Discriminator conditioning significantly improved perceptual quality

S2 vs. (S4, S5, S6)

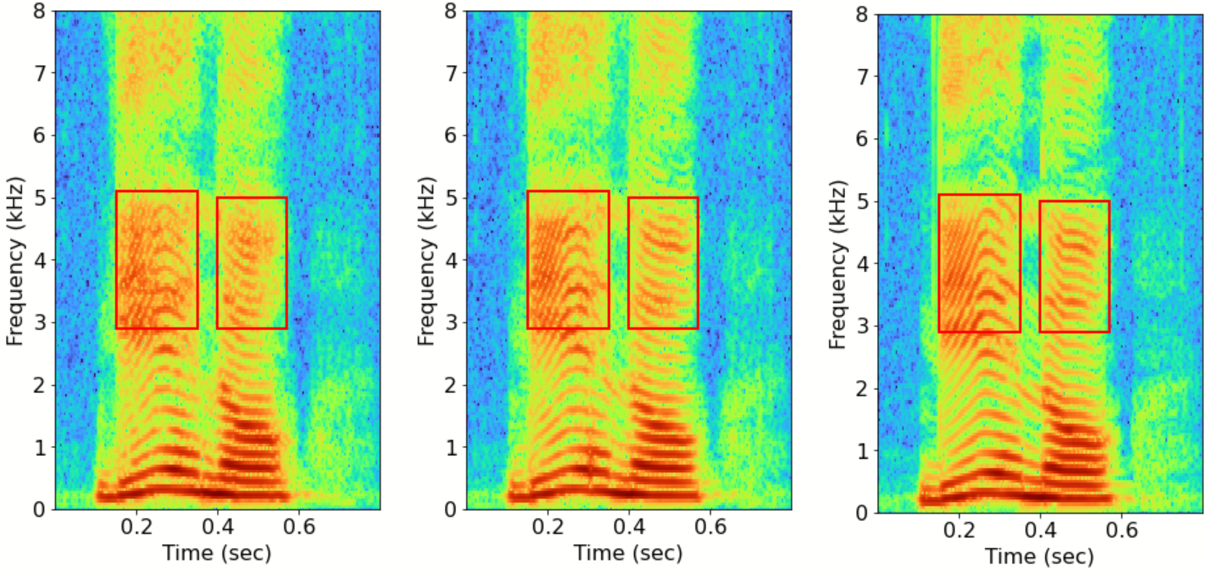
(Intentionally) misconfigured discriminators degraded performance

S2 vs. S7

Property designed voicing-aware discriminator worked best.

R. Yamamoto, *et al.*, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.

Comparison of spectrograms

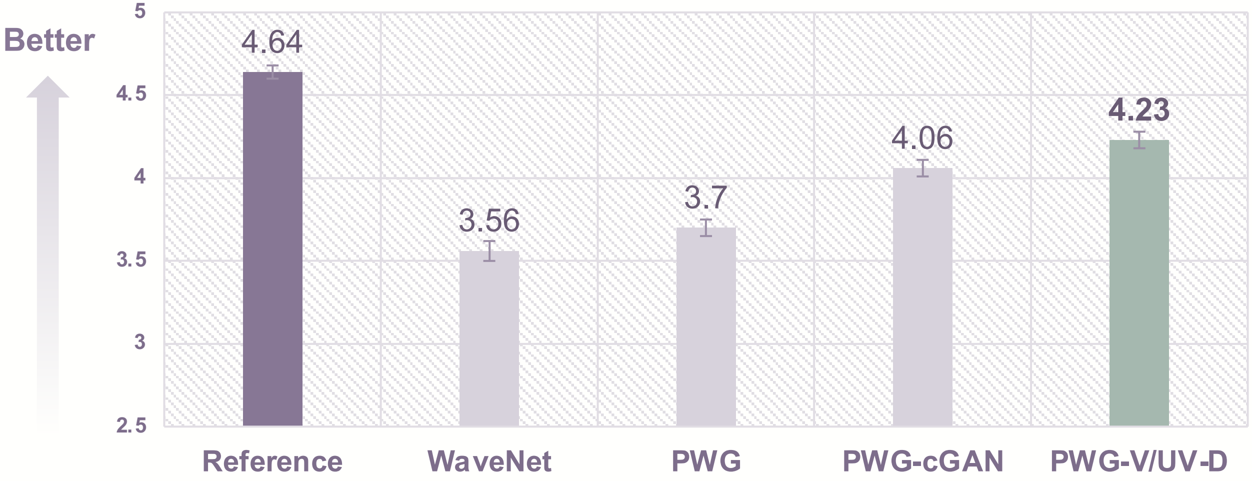


PWG **PWG-V/UV-D** **Reference**

PWG-V/UV-D can produce spectral harmonics more accurately.

R. Yamamoto, *et al.*, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.

MOS listening test results on text-to-speech



NOTE: MOS in the figure was averaged among four speakers. See per-speaker MOS in our paper.
NOTE: WaveNet could be improved by noise-shaping technique.

Female

Male

R. Yamamoto, *et al.*, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.

Summary

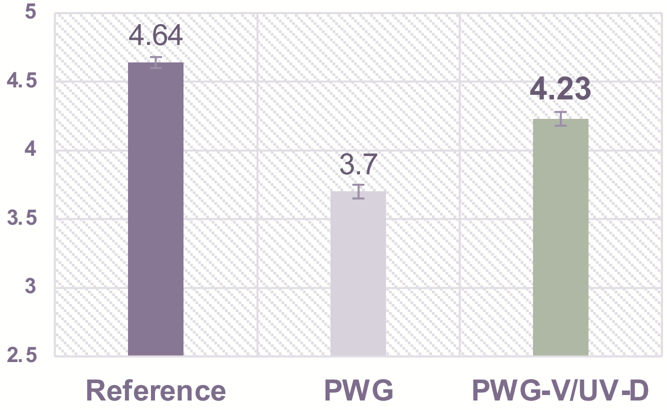
Goal

Better perceptual quality by improving PWG's *discriminator*

Proposed method

Voicing-aware discriminator: separate discriminators for voiced and unvoiced segments.

Results



Demo samples

R. Yamamoto, *et al.*, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *Proc. ICASSP*, 2021.