# Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram

***Ryuichi Yamamoto***[1], *Eunwoo Song*[2], *Jae-min Kim*[2]

[1]LINE Corp., Japan

[2]NAVER Corp., South Korea

# Raw waveform generation: Autoregressive (AR) vs. non-AR

## Autoregressive models

☺ High-fidelity speech generation (e.g., **WaveNet** [1])

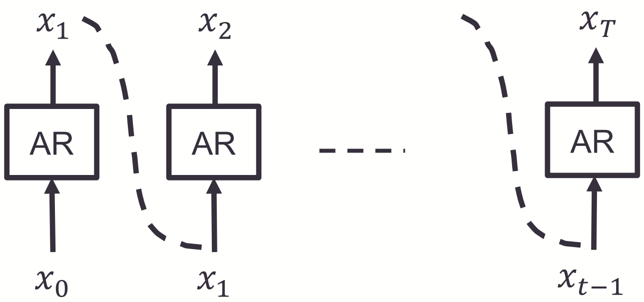
☹ Generation is too **slow**

## Non-autoregressive models

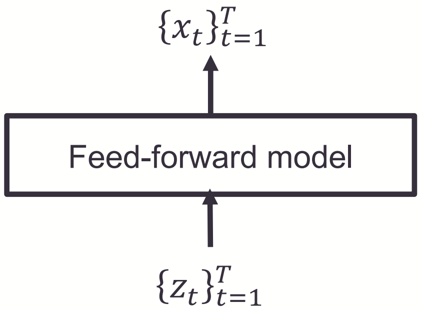
Teacher-student-based methods (**Parallel WaveNet** [2], **ClariNet** [3])

☺ Real-time generation

☹ Complicated two-stage training using probability density distillation




Autoregressive generation (O(T) in time)




Non-autoregressive generation (O(1) in time)


[1] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.
[2] A. van den Oord, *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. ICML*, 2018.
[3] W. Ping, *et al.*, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICLR*, 2019.

# Our approach: GANs for waveform generation

Parallel WaveGAN (Parallel inference + WaveNet + GAN)

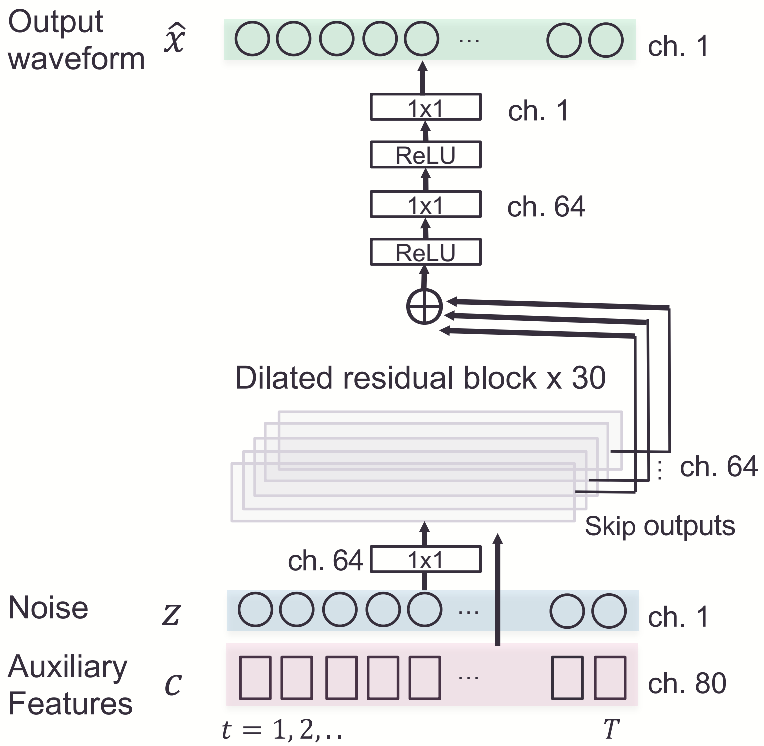
- **Distillation-free:** a distillation-free fast waveform generation, combining multi-resolution STFT loss and adversarial loss.
- **Fast:** Training and inference speed become 4.82 / 1.96 times faster than the conventional parallel WaveNet (i.e. ClariNet).
- **High-quality:** Our model achieves 4.16 MOS (in Transformer-based TTS) that is competitive to the best distillation-based ClariNet.

*GAN-based methods can be good alternatives to distillation based methods.*

STFT: Short-time Fourier transform
MOS: Mean-opinion score

Yamamoto et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020.

# Parallel WaveGAN: WaveNet-based generator




1x1: 1x1 convolution

## Architecture

Generator architecture is almost the same as **WaveNet [1]**

## Conditional waveform generation

80-dim mel-spectrogram as auxiliary features

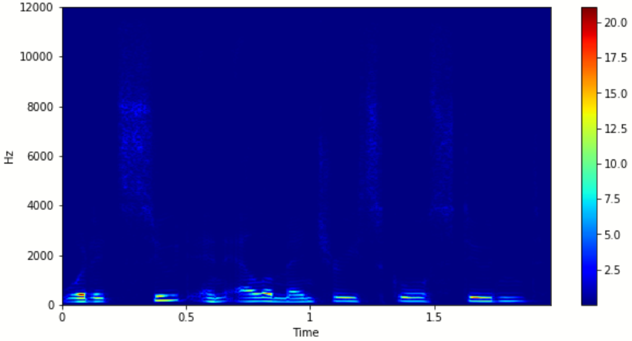
Model comparison between WaveNet and ours

| | **WaveNet** | **Parallel WaveGAN** |
|---|---|---|
| Input | Previous samples | Random noise for all time steps |
| Output | Probability distribution | Raw waveform samples |
| Convolution | Causal conv. | Non-causal conv. |

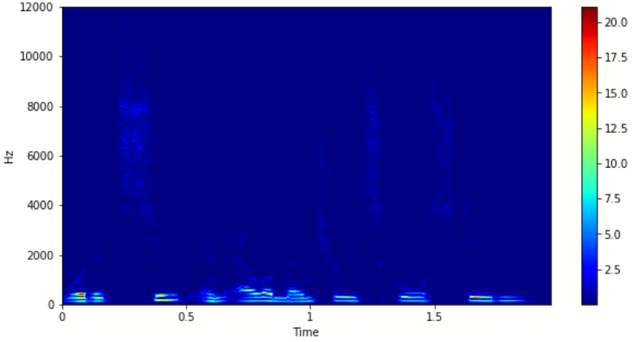

[1] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.
Yamamoto et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020.

# STFT loss: Spectral convergence (SC) [4]

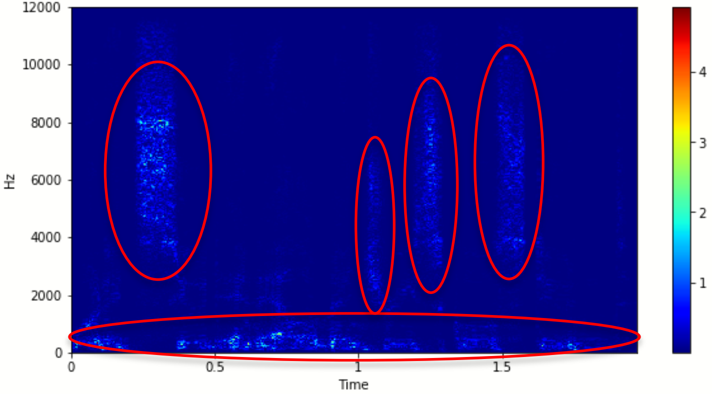
$|\mathrm{STFT}(x)|$




$|\mathrm{STFT}(\hat{x})|$




$||\mathrm{STFT}(x)| - |\mathrm{STFT}(\hat{x})||$




SC penalizes large amplitude components

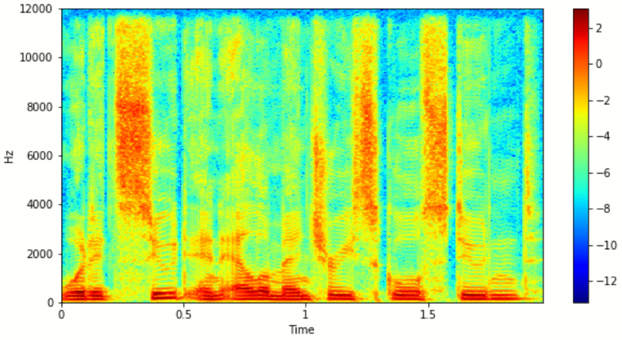
$$L_{sc} = \frac{\| |\mathrm{STFT}(x)| - |\mathrm{STFT}(\hat{x})| \|_F}{\| |\mathrm{STFT}(x)| \|_F}$$


[4] S. O. Arik, et al, “Fast Spectrogram Inversion using Multi-head Convolutional Neural Networks,” IEEE Signal Procees. Letters, 2019.

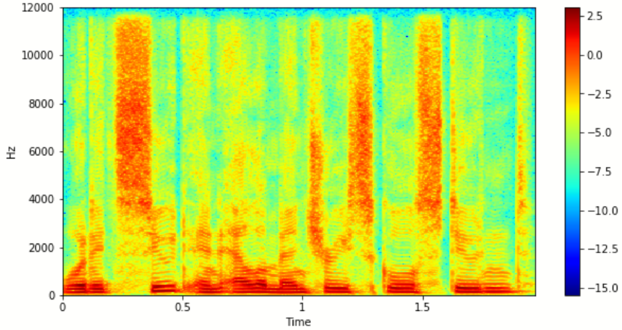
Yamamoto et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020.

# STFT loss: Log-scale STFT magnitude loss [4]

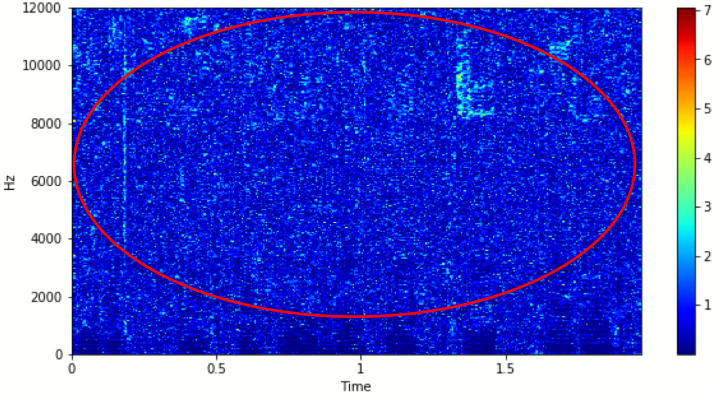
$\log|\mathrm{STFT}(x)|$




$\log|\mathrm{STFT}(\hat{x})|$




$\big|\log|\mathrm{STFT}(x)| - \log|\mathrm{STFT}(\hat{x})|\big|$




Log STFT loss penalizes small amplitude components

$$L_{\mathrm{mag}} = \frac{1}{N}\big\|\log|\mathrm{STFT}(x)| - \log|\mathrm{STFT}(\hat{x})|\big\|_1$$

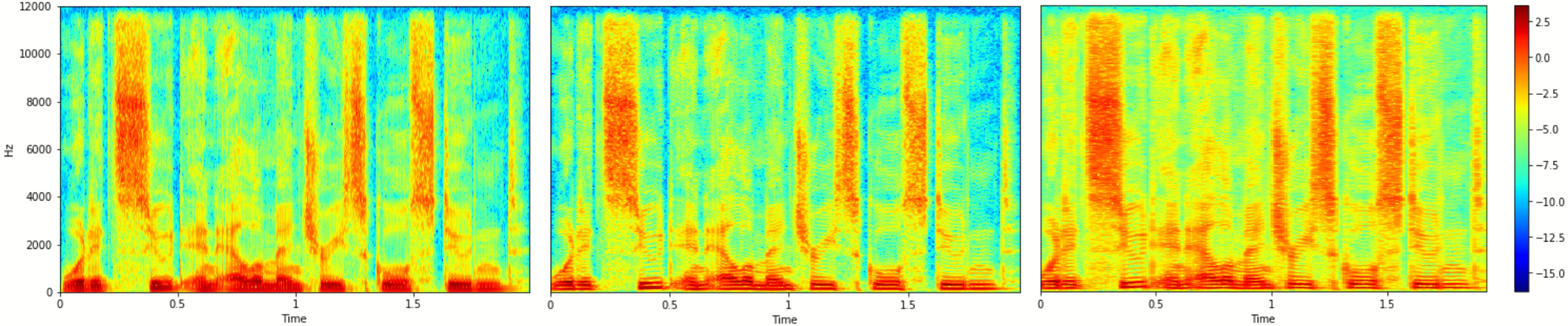
N: number of elements in the STFT magnitude

[4] S. O. Arik, et al, “Fast Spectrogram Inversion using Multi-head Convolutional Neural Networks,” IEEE Signal Procees. Letters, 2019.

# Multi-resolution STFT loss

## FFT size / window size / shift

512 / 240 / 50 | 1024 / 600 / 120 | 2048 / 1200 / 240



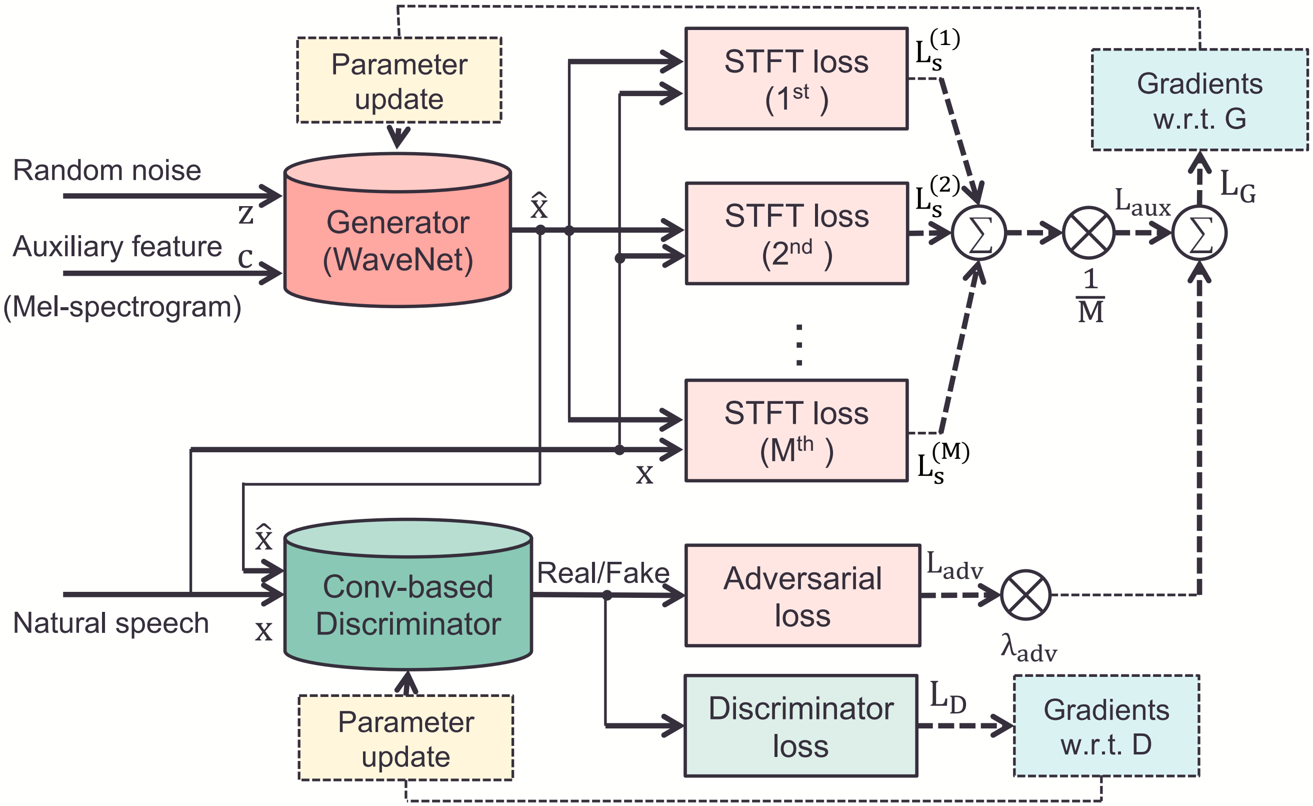

**Higher temporal resolution** | Balanced | **Higher frequency resolution**

$$L_{\mathrm{aux}}(G) = \frac{1}{M}\sum_{m=1}^{M} L_{\mathrm{s}}^{(m)}(G)$$

$$L_{\mathrm{s}}(G) = \mathbb{E}_{z\sim p(z),x\sim p_{\mathrm{data}}}\left[L_{\mathrm{sc}}(x,\hat{x}) + L_{\mathrm{mag}}(x,\hat{x})\right]$$

M: number of STFT losses

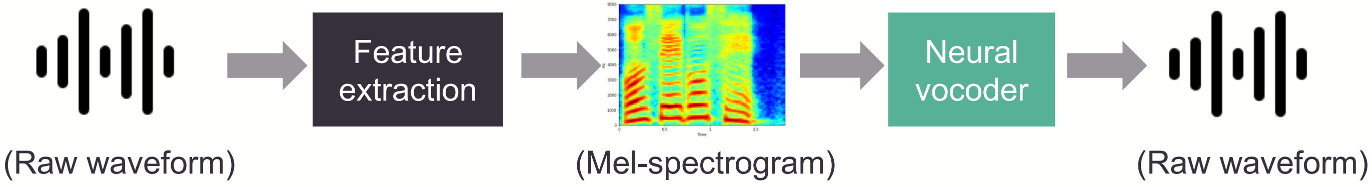
# Parallel WaveGAN: Training overview



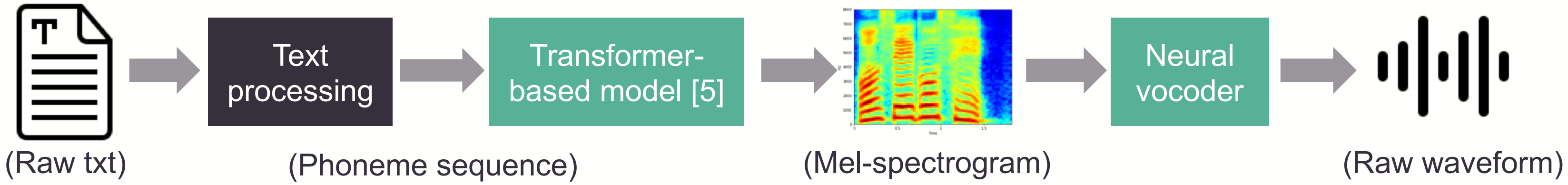

Yamamoto et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020.

# Experiments

## 1) Analysis/synthesis




## 2) Text-to-speech




[5] N.Li,et al, “Neural speech synthesis with Transformer network,” in Proc. AAAI, 2019.

Yamamoto et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020.

# Experimental conditions

## Data & features

| Recordings | Size (training / validation / test) |
|---|---|
| 24 kHz /16 bit, female professional Japanese speaker | 11,449 (23 hours) / 250 / 250 |

| Auxiliary features | Frame shift | Frame length | Frequency range |
|---|---|---|---|
| 80-dim log-melspectrogram | 12.5 ms | 50 ms | 70 - 8000 Hz |

## Vocoder model comparison

- Single Gaussian WaveNet [1,3]
- ClariNet (single / three STFT losses) [3]
- ClariNet-GAN (single / three STFT losses) [6]
- Parallel WaveGAN (single / three STFT losses)

## Listening tests

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

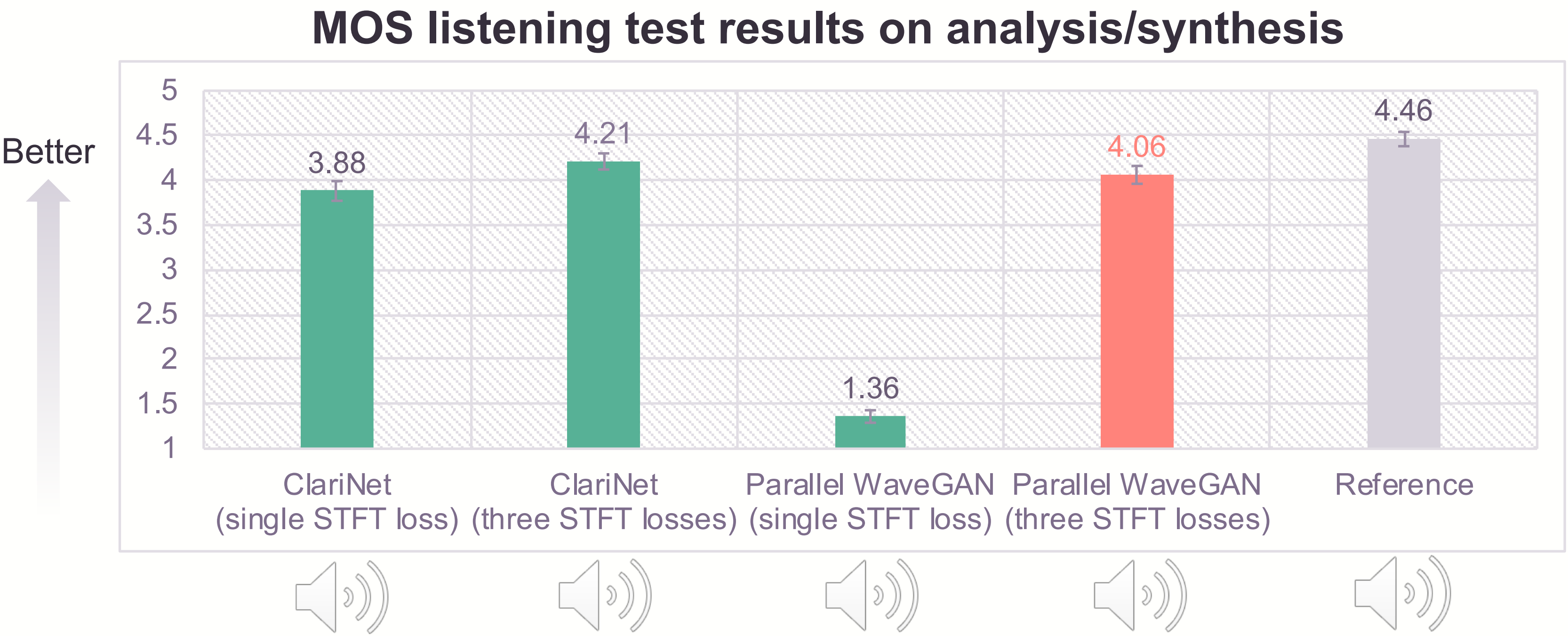
18 native Japanese speakers / 20 random utterances for each model

[1] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[3] W. Ping, *et al.*, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICLR*, 2019.

[6] R. Yamamoto *et al.*, "Probability density distillation with generative adversarial networks for high-quality parallel waveform generation," in Proc. INTER-SPEECH, 2019.

# Analysis/synthesis: Effects of multi-resolution STFT loss




Using multi-resolution STFT loss largely improved perceptual quality for both ClariNet and Parallel WaveGAN.

Yamamoto et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020.

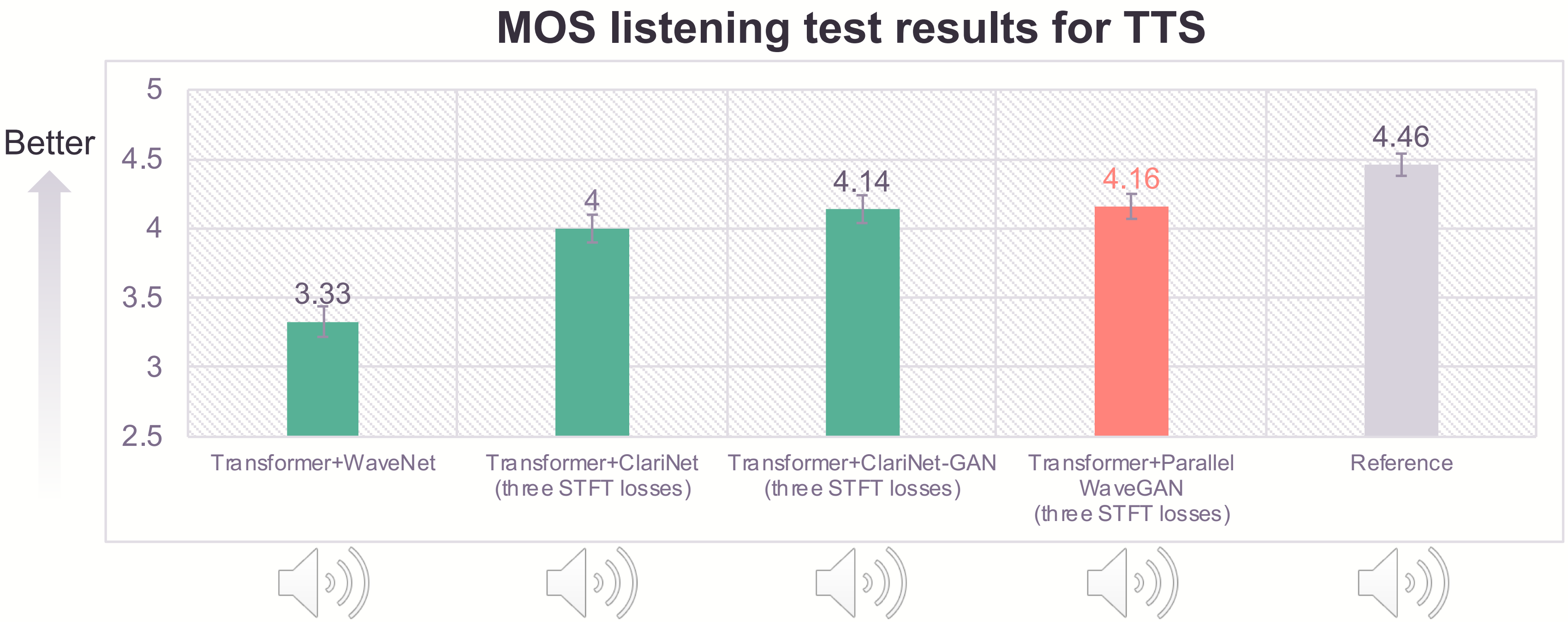
# Training/inference time and model size comparison

| Model | Training time (in days) | Inference speed (k times faster than real-time) | Number of parameters (in millions) |
|---|---|---|---|
| WaveNet | 7.4 | 0.0032 | 3.81 |
| ClariNet | 12.7 | 14.62 | 2.78 |
| **Parallel WaveGAN (ours)** | **2.8** | **28.68** | **1.44** |
| | Lower is better | Higher is better | Lower is better |

All training was conducted on a server with two NVIDIA Tesla V100 GPUs.
All inference test was conduced on a server with a single NVIDIA Tesla V100 GPU.

Yamamoto et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020.

# Text-to-speech: Perceptual quality evaluation




Our model achieved 4.16 MOS competitive to the best distillation-based ClariNet.

Yamamoto et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020.

# Conclusion

## Goal

Fast, high-quality and simple waveform generation for text-to-speech (TTS)

## Proposed method

Parallel WaveGAN, a distillation-free fast waveform generation, combining multi-resolution STFT loss and adversarial loss.

## Results

Comparative perceptual quality (MOS 4.16 in Transformer-based TTS) to the best distillation-based method while improving inference and training speed.

*Take-home message: GAN-based methods can be good alternatives to distillation based methods.*

Yamamoto et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020.

# Acknowledgements

The work was supported by Clova Voice, NAVER Corp., Seongnam, Korea. The authors would like to thank Adrian Kim, Jung-Woo Ha, Muhammad Ferjad Naeem, and Xiaodong Gu at NAVER Corp., Seongnam, Korea, for their support.

## Demos

https://r9y9.github.io/demos/projects/icassp2020/

## Any questions?

ryuichi.yamamoto@linecorp.com

LINE zryuichi