# Parallel waveform synthesis

Eunwoo Song / Naver Clova

# Who am I?

## Education

- B.S., E.E., Yonsei Univ., Seoul, Korea (Aug 2010)
- Combined M.S. and Ph.D., EE., Yonsei Univ., Seoul, Korea (Feb 2019)

## Work experience

- NAVER Corp., Seongnam, Korea
  - Senior Research Scientist (Mar 2017 - present)
  - DNN TTS Team Lead, Clova Voice
- Seoul National Univ., Seoul, Korea
  - Adjunct professor, Artificial Intelligence Institute (Aug 2022 - present)

# Who am I?

Education

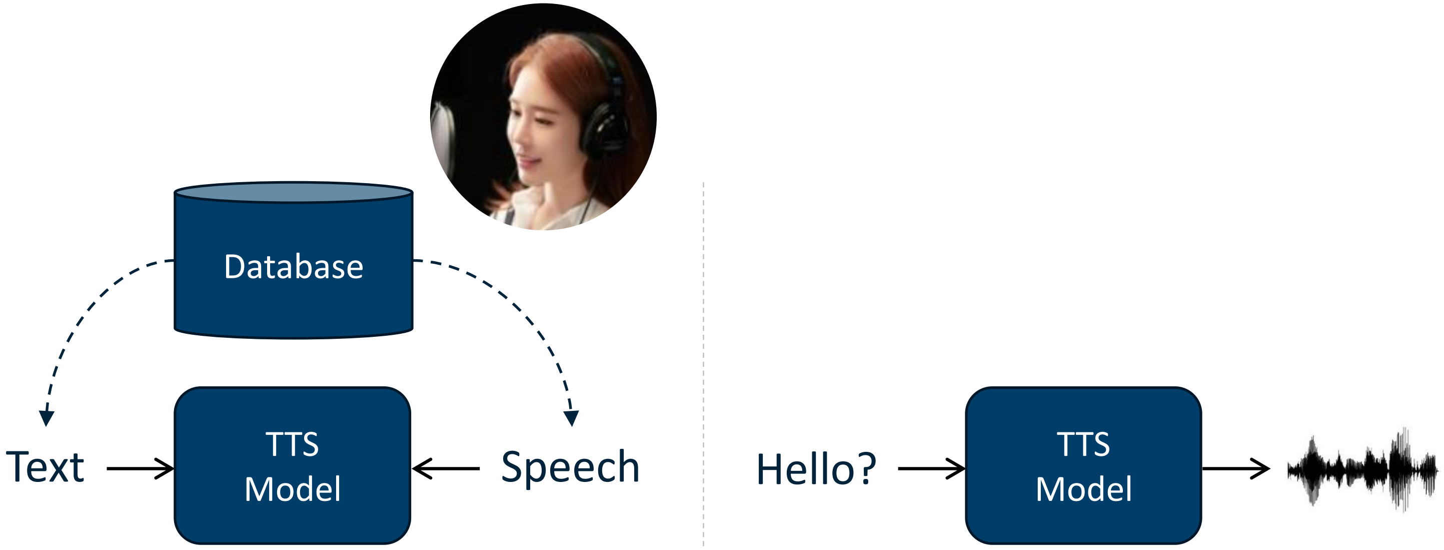
- B.S., E.E., Yonsei Univ., Seoul, Korea (Aug 2010)
- Combined M.S. and Ph.D., EE., Yonsei Univ., Seoul, Korea (Feb 2019)

Work experience

- NAVER Corp., Seongnam, Korea
  - Senior Research Scientist (Mar 2017 - present)
  - DNN TTS Team Lead, Clova Voice
- Seoul National Univ., Seoul, Korea
  - Adjunct professor, Artificial Intelligence Institute (Aug 2022 - present)

# Introduction

## Deep learning-based TTS system




Training

Inference

**Human-like voice quality**

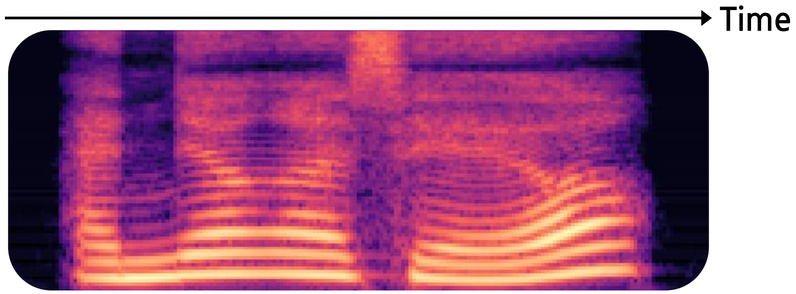
# Introduction

## Deep learning-based TTS system


**DNN TTS = Acoustic model + Vocoding model**

# Introduction

## Deep learning-based TTS system



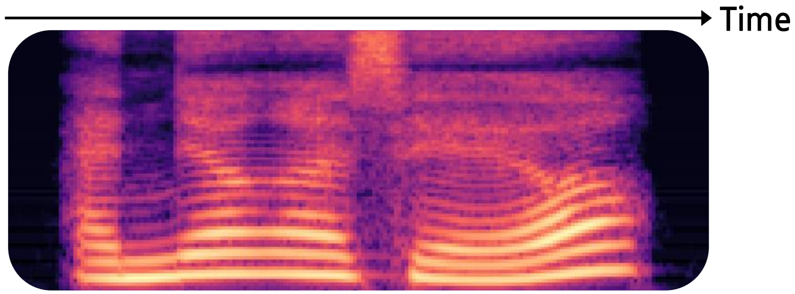


Estimating acoustic parameters from text inputs

**DNN TTS = Acoustic model + Vocoding model**

# Introduction

## Deep learning-based TTS system



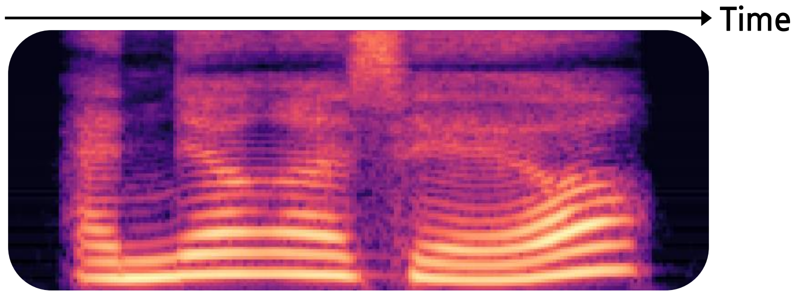


Estimating speech signals from acoustic parameters

**DNN TTS = Acoustic model + Vocoding model**

# Introduction

## Deep learning-based TTS system





Estimating speech signals from acoustic parameters

**DNN TTS = Acoustic model + Vocoding model**

# Introduction

## PARALLEL WAVEGAN: A FAST WAVEFORM GENERATION MODEL BASED ON GENERATIVE ADVERSARIAL NETWORKS WITH MULTI-RESOLUTION SPECTROGRAM

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

[1]LINE Corp., Tokyo, Japan.
[2]NAVER Corp., Seongnam, Korea

**ABSTRACT**

We propose Parallel WaveGAN, a distillation-free, fast, and small-footprint waveform generation method using a generative adversarial network. In the proposed method, a non-autoregressive WaveNet is trained by jointly optimizing multi-resolution spectrogram and adversarial loss functions, which can effectively capture the time-frequency distribution of the realistic speech waveform. As our method does not require density distillation used in the conventional teacher-student framework, the entire model can be easily trained. Furthermore, our model is able to generate high-fidelity speech even with its compact architecture. In particular, the proposed Parallel WaveGAN has only 1.44 M parameters and can generate 24 kHz speech waveform 28.68 times faster than real-time on a single GPU environment. Perceptual listening test results verify that our proposed method achieves 4.16 mean opinion score within a Transformer-based text-to-speech framework, which is comparative to the best distillation-based Parallel WaveNet system.

# Parallel waveform synthesis

Vocoding models: Overview

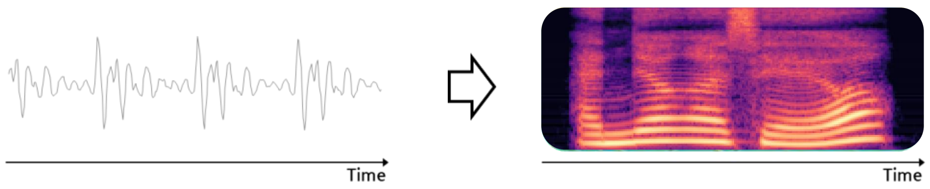
# Vocoding models: Overview

## Estimating speech signals from acoustic parameters

Acoustic Parameters → **Vocoding Model** → 

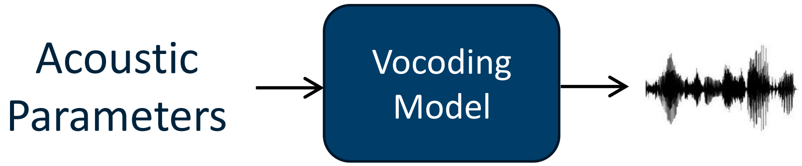
Acoustic parameters..?

Representing speech characteristics
such as F0, spectrum, v/uv …



# Vocoding models: Overview

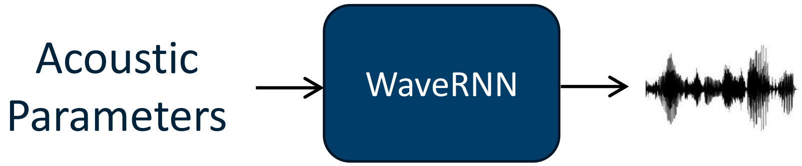
## Estimating speech signals from acoustic parameters




What is the main model?

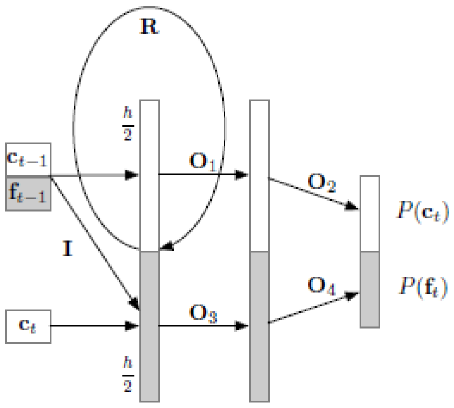
# Vocoding models: Overview

## Estimating speech signals from acoustic parameters




What is the main model?

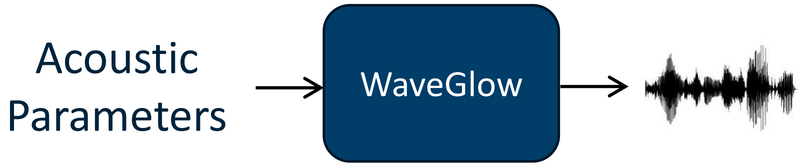
WaveRNN based on the RNN model




N. Kalchbrenner, et al., “Efficient neural audio synthesis,” arXiv:1802.08435, 2018.

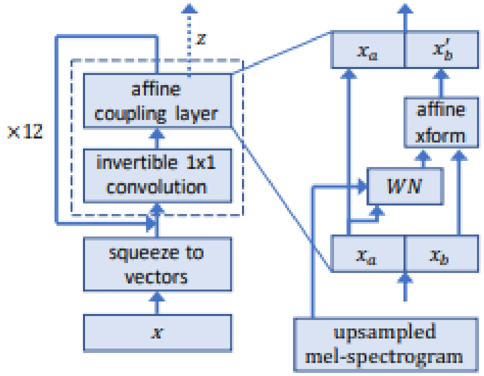
# Vocoding models: Overview

## Estimating speech signals from acoustic parameters




What is the main model?

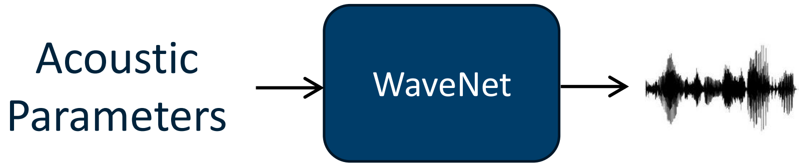
WaveGlow based on the Flow model




R. Prenger, et al., "WaveGlow: A flow-based generative network for speech synthesis." in Proc. ICASSP, 2019.

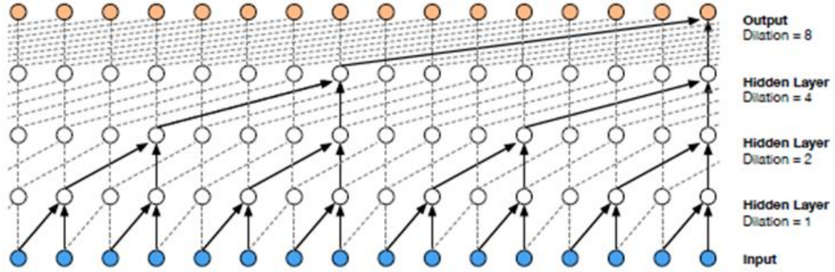
# Vocoding models: Overview

## Estimating speech signals from acoustic parameters




What is the main model?

WaveNet based on the CNN model



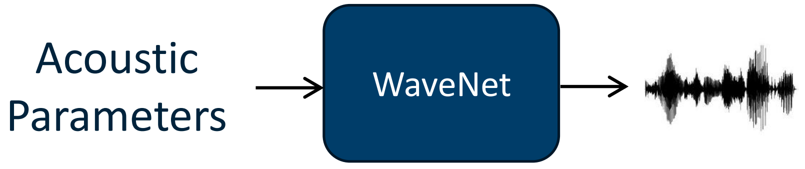

$$p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^{T} p(x_t|x_1, \cdots, x_{t-1}, \mathbf{h})$$

$$\mathbf{z} = \tanh(W_{f,k} * \mathbf{x} + \mathbf{V}_{f,k}^T\mathbf{h}) \odot \delta(W_{g,k} * \mathbf{x} + \mathbf{V}_{g,k}^T\mathbf{h})$$

A. Van Den Oord, et al., “WaveNet: A generative model for raw audio,” CoRR abs/1609.03499, 2016.

# Vocoding models: Overview

## Estimating speech signals from acoustic parameters




What is the main model?

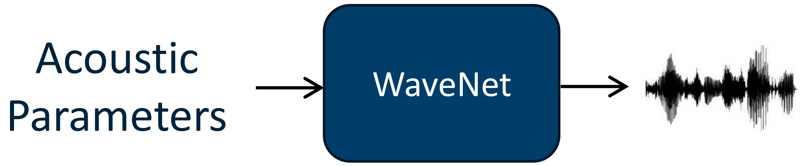
WaveNet based on the CNN model

Estimating the current sample from the previous samples
We define this method as autoregressive vocoding model

A. Van Den Oord, et al., “WaveNet: A generative model for raw audio,” CoRR abs/1609.03499, 2016.

# Vocoding models: Overview

## Estimating speech signals from acoustic parameters




What is the main model?

WaveNet based on the CNN model

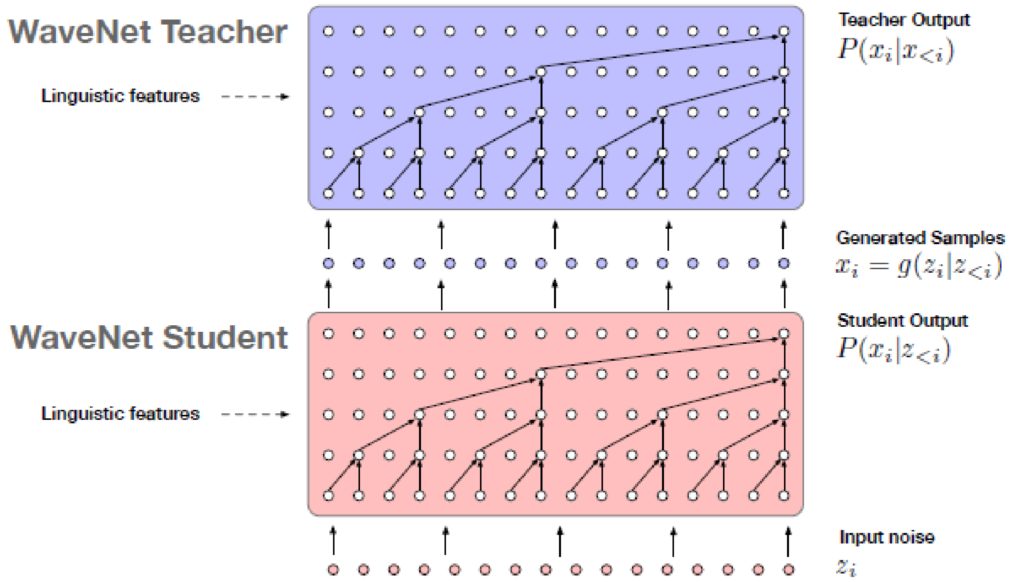
Estimating the current sample from the previous samples
We define this method as autoregressive vocoding model

WaveNet generates high-quality synthetic speech
However, it takes about 5 minutes to generate 1 sec audio

A. Van Den Oord, et al., “WaveNet: A generative model for raw audio,” CoRR abs/1609.03499, 2016.

# Vocoding models: Overview

## Estimating speech signals from acoustic parameters

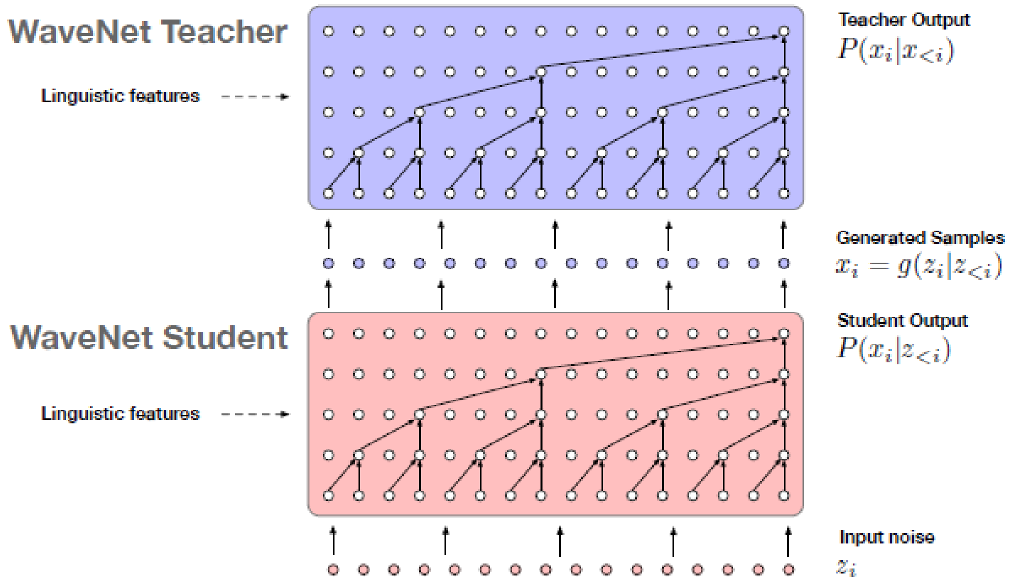



One of the alternative method to address WaveNet's slow inference speed is the non-autoregressive Parallel WaveNet

A. van den Oord, et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in Proc. ICML, 2018.

# Vocoding models: Overview

## Estimating speech signals from acoustic parameters

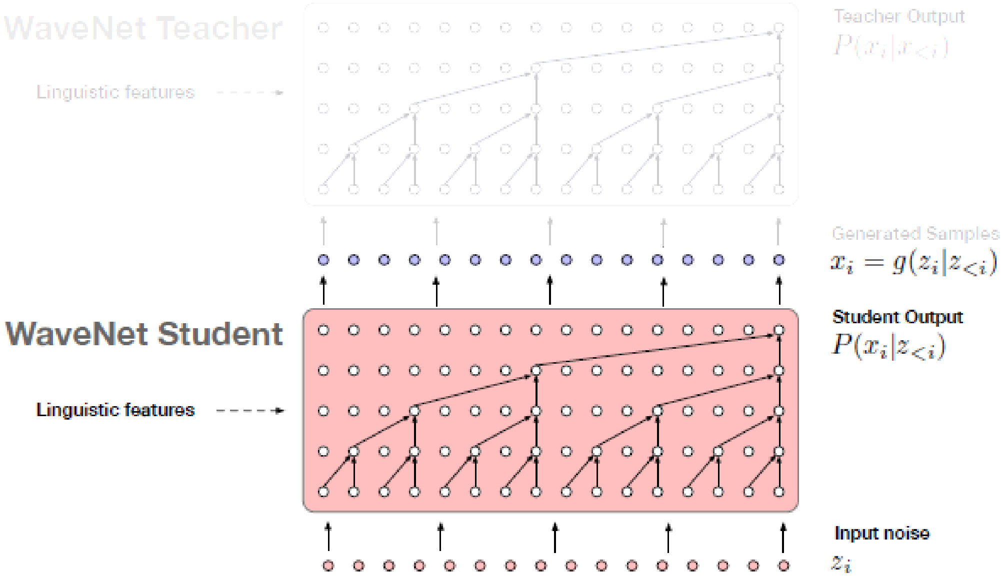



Non-autoregressive Parallel WaveNet (=student) is trained to learn the distribution of the autoregressive WaveNet (=teachure)

A. van den Oord, et al., “Parallel WaveNet: Fast high-fidelity speech synthesis,” in Proc. ICML, 2018.

# Vocoding models: Overview

## Estimating speech signals from acoustic parameters

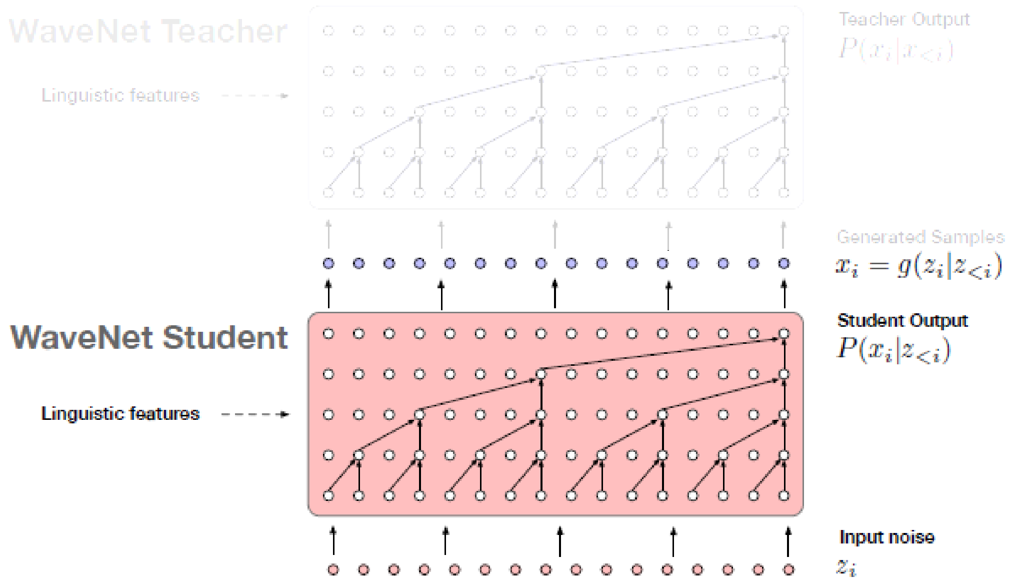



Non-autoregressive Parallel WaveNet doesn't require the previous samples
Its inference speed in unlimited
(it takes about 0.02 sec to generate 1 sec audio)

A. van den Oord, et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in Proc. ICML, 2018.

# Vocoding models: Overview

## Estimating speech signals from acoustic parameters




There remain problems in the difficult training method…

A. van den Oord, et al., “Parallel WaveNet: Fast high-fidelity speech synthesis,” in Proc. ICML, 2018.

# Parallel waveform synthesis

Vocoding models: Parallel WaveGAN

# Vocoding models: Parallel WaveGAN

PARALLEL WAVEGAN: A FAST WAVEFORM GENERATION MODEL BASED ON GENERATIVE ADVERSARIAL NETWORKS WITH MULTI-RESOLUTION SPECTROGRAM

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

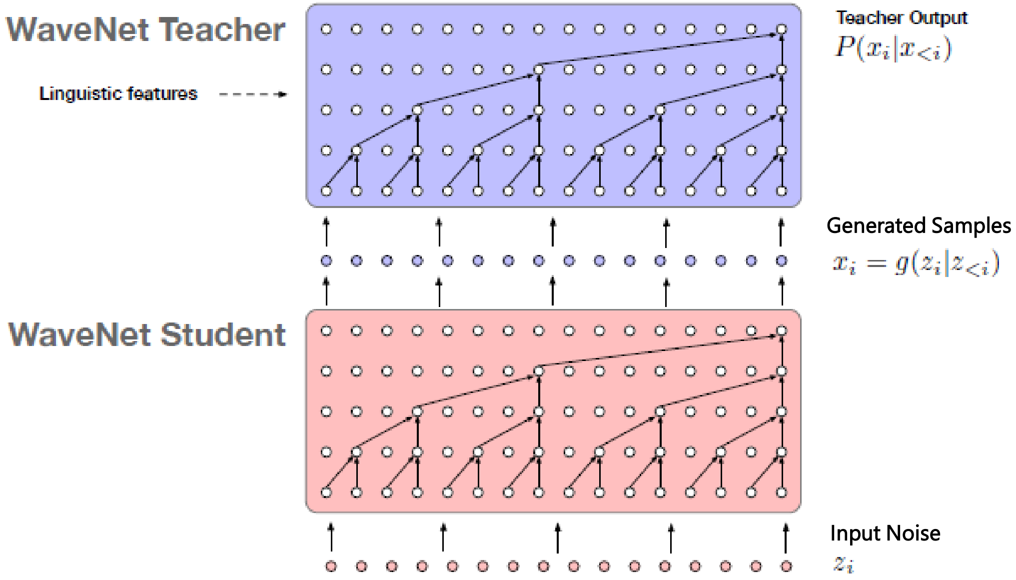
[1]LINE Corp., Tokyo, Japan.
[2]NAVER Corp., Seongnam, Korea

**ABSTRACT**

We propose Parallel WaveGAN, a distillation-free, fast, and small-footprint waveform generation method using a generative adversarial network. In the proposed method, a non-autoregressive WaveNet is trained by jointly optimizing multi-resolution spectrogram and adversarial loss functions, which can effectively capture the time-frequency distribution of the realistic speech waveform. As our method does not require density distillation used in the conventional teacher-student framework, the entire model can be easily trained. Furthermore, our model is able to generate high-fidelity speech even with its compact architecture. In particular, the proposed Parallel WaveGAN has only 1.44 M parameters and can generate 24 kHz speech waveform 28.68 times faster than real-time on a single GPU environment. Perceptual listening test results verify that our proposed method achieves 4.16 mean opinion score within a Transformer-based text-to-speech framework, which is comparative to the best distillation-based Parallel WaveNet system.

# Vocoding models: Parallel WaveGAN

1. Removed the teacher-student distillation process

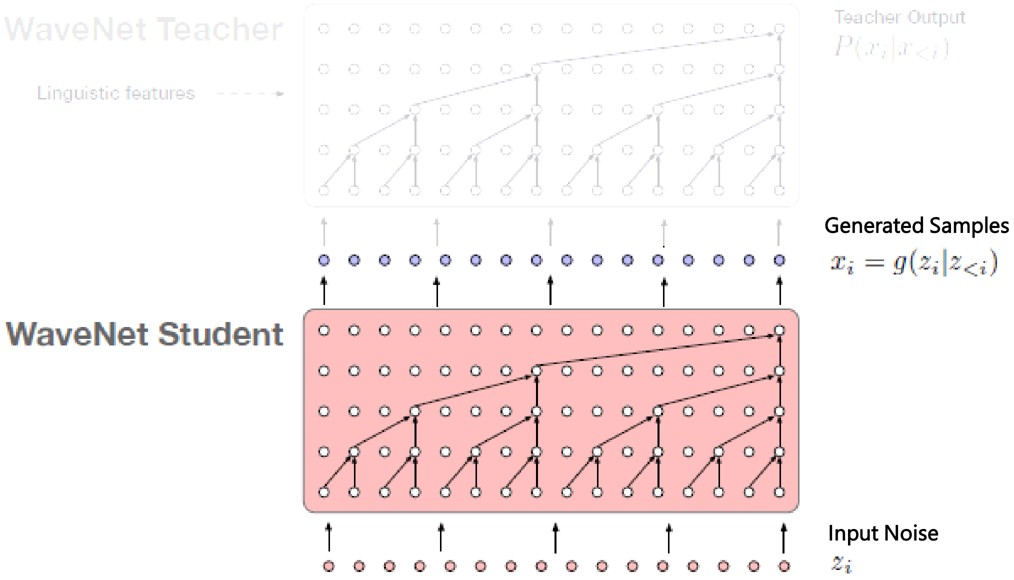



R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

1. Removed the teacher-student distillation process

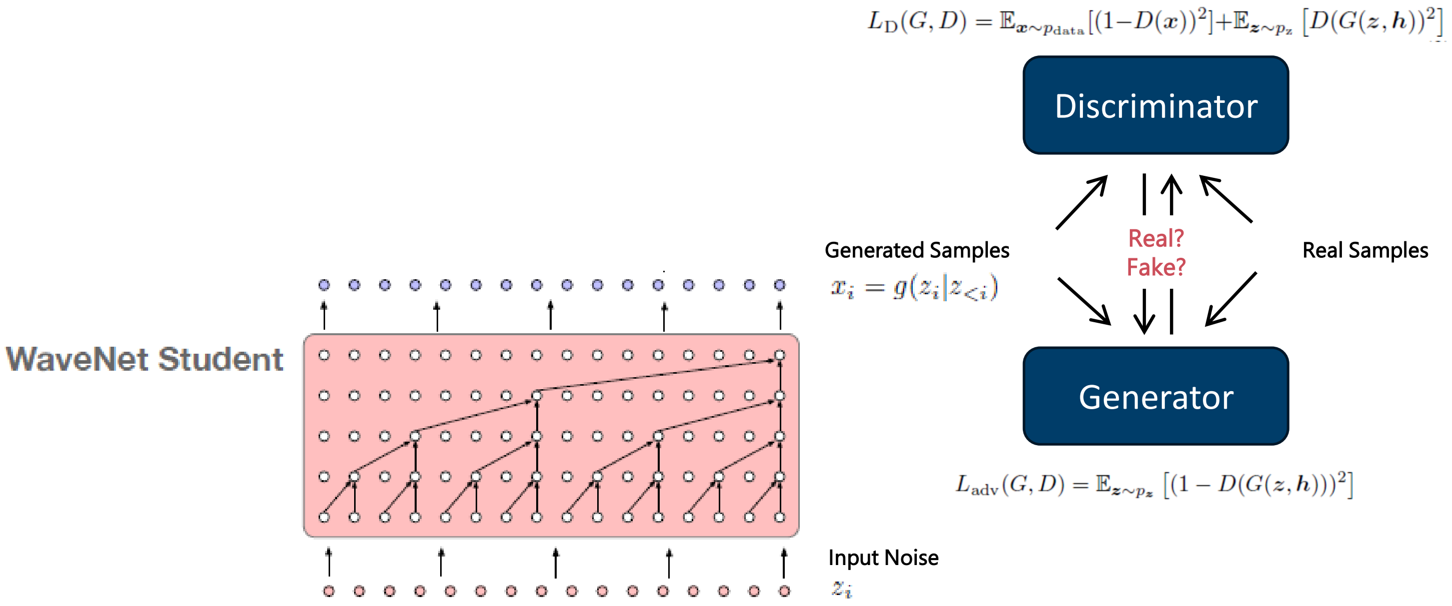
→ Entire model can be "easily" trained




R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

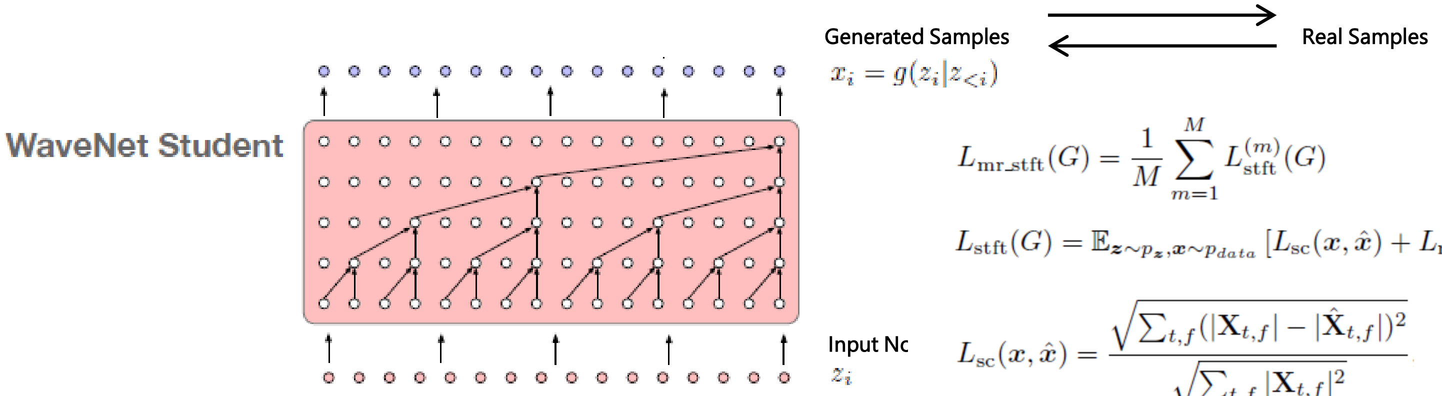
1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method




R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method
3. Further improved its quality by introducing the multi-resolution STFT loss




$$L_{\mathrm{mr\_stft}}(G) = \frac{1}{M}\sum_{m=1}^{M} L_{\mathrm{stft}}^{(m)}(G)$$

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

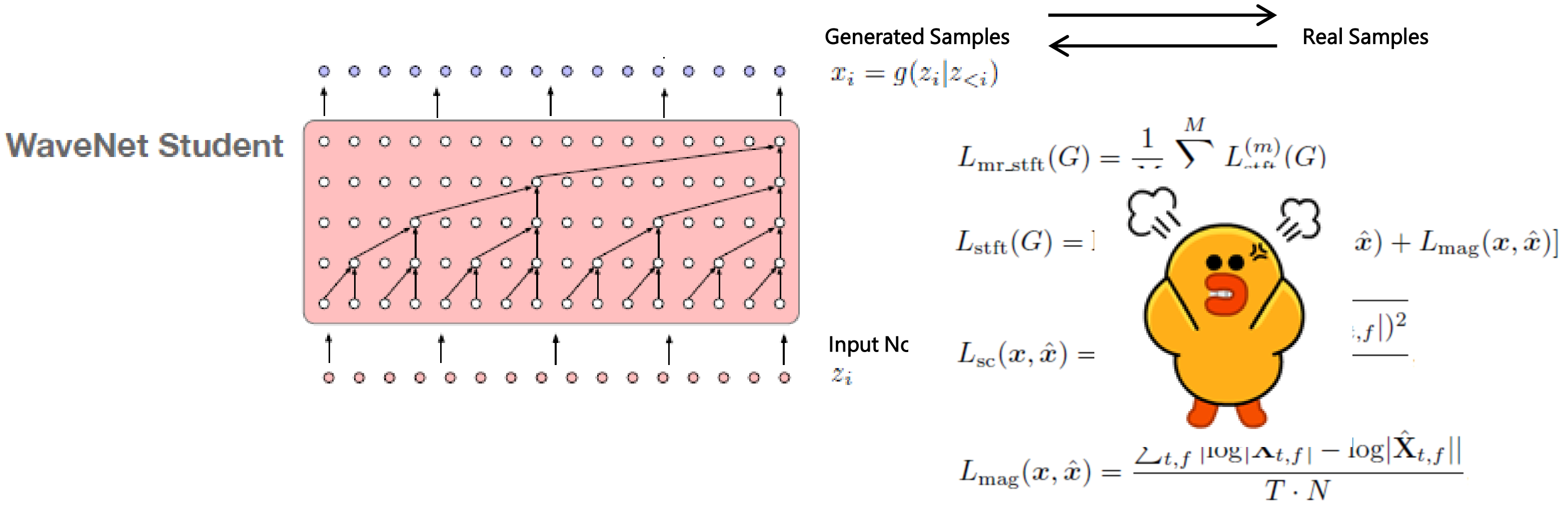
$$L_{\mathrm{sc}}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sqrt{\sum_{t,f}(|\mathbf{X}_{t,f}| - |\hat{\mathbf{X}}_{t,f}|)^2}}{\sqrt{\sum_{t,f}|\mathbf{X}_{t,f}|^2}}$$

$$L_{\mathrm{mag}}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sum_{t,f}\left|\log|\mathbf{X}_{t,f}| - \log|\hat{\mathbf{X}}_{t,f}|\right|}{T\cdot N}$$


R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method
3. Further improved its quality by introducing the multi-resolution STFT loss



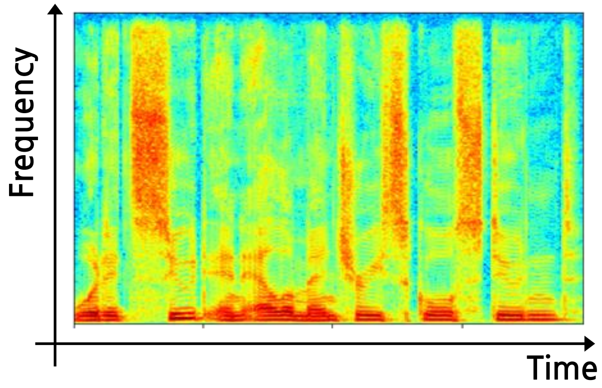

R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method
3. Further improved its quality by introducing the multi-resolution STFT loss

STFT (short-time Fourier transform)?

Time-frequency representation of speech signal



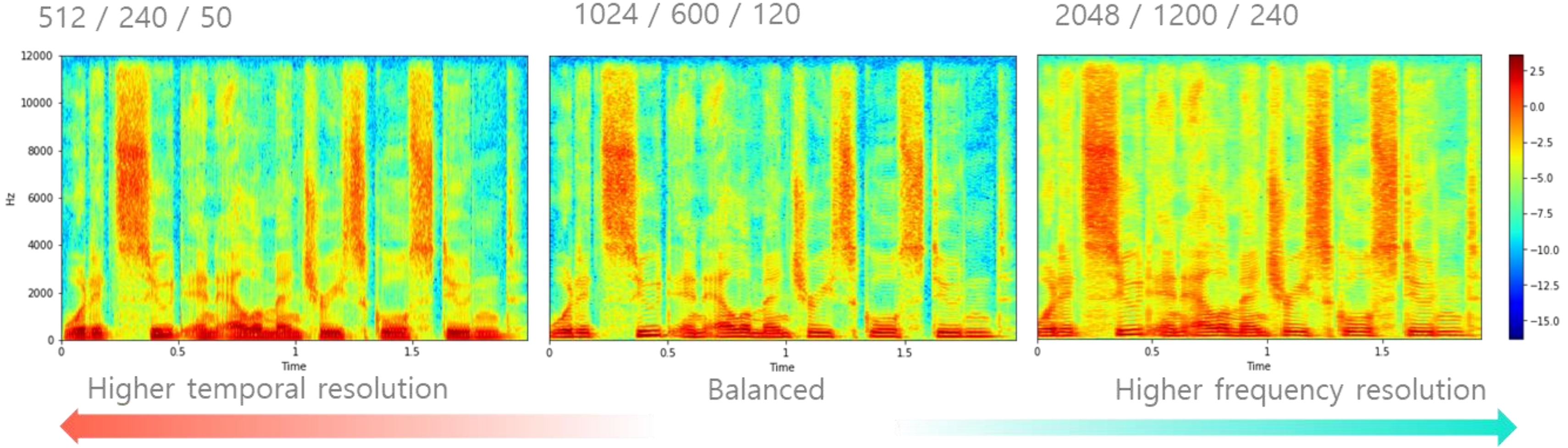

R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method
3. Further improved its quality by introducing the multi-resolution STFT loss

STFT is calculated in different T/F resolutions

FFT size / window size / shift



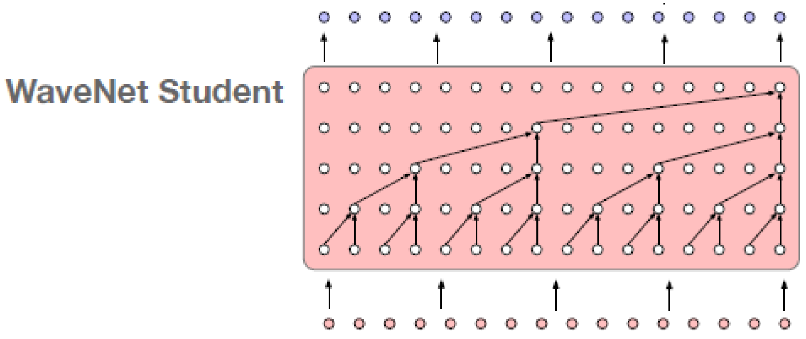

R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method
3. Further improved its quality by introducing the multi-resolution STFT loss

STFT is calculated in different T/F resolutions

There are two loss functions




Generated Samples ⇄ Real Samples

$x_i = g(z_i|z_{<i})$

Input Nc

$z_i$

$$L_{\mathrm{mr\_stft}}(G) = \frac{1}{M}\sum_{m=1}^{M} L_{\mathrm{stft}}^{(m)}(G)$$

$$L_{\mathrm{stft}}(G) = \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}},\boldsymbol{x}\sim p_{data}} \boxed{[L_{\mathrm{sc}}(\boldsymbol{x},\hat{\boldsymbol{x}}) + L_{\mathrm{mag}}(\boldsymbol{x},\hat{\boldsymbol{x}})]}$$

$$L_{\mathrm{sc}}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sqrt{\sum_{t,f}(|\mathbf{X}_{t,f}| - |\hat{\mathbf{X}}_{t,f}|)^2}}{\sqrt{\sum_{t,f}|\mathbf{X}_{t,f}|^2}}$$

$$L_{\mathrm{mag}}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sum_{t,f}\left|\log|\mathbf{X}_{t,f}| - \log|\hat{\mathbf{X}}_{t,f}|\right|}{T\cdot N}$$


R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

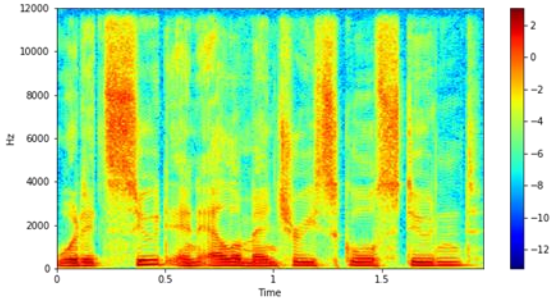
# Vocoding models: Parallel WaveGAN

1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method
3. Further improved its quality by introducing the multi-resolution STFT loss

STFT is calculated in different T/F resolutions

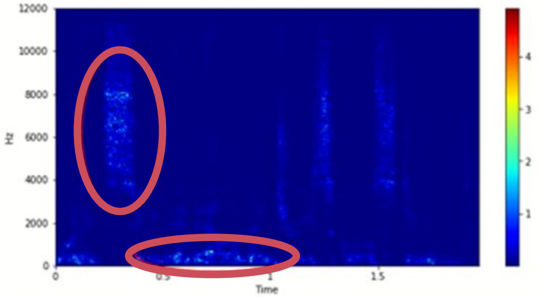
There are two loss functions

One penalizes large energy components



Real Spectrogram

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



SC penalizes large amplitude components

$$L_{\mathrm{sc}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\sqrt{\sum_{t,f}(|\mathbf{X}_{t,f}| - |\hat{\mathbf{X}}_{t,f}|)^2}}{\sqrt{\sum_{t,f}|\mathbf{X}_{t,f}|^2}}$$

R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

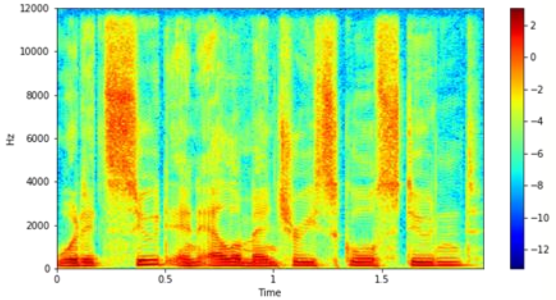
1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method
3. Further improved its quality by introducing the multi-resolution STFT loss

STFT is calculated in different T/F resolutions

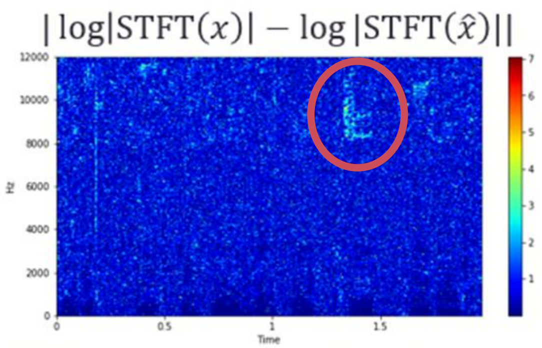
There are two loss functions

One penalizes large energy components

The other penalizes small energy components




Real Spectrogram




Log STFT loss penalizes small amplitude components

$$L_{\mathrm{mag}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\sum_{t,f} |\log|\mathbf{X}_{t,f}| - \log|\hat{\mathbf{X}}_{t,f}||}{T \cdot N}$$

R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

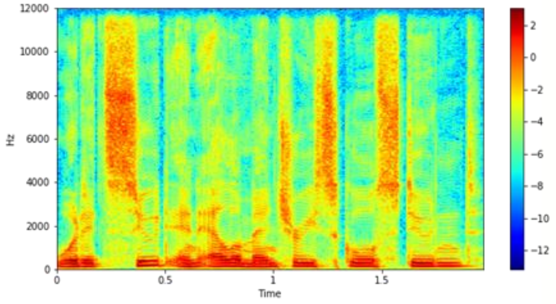
1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method
3. Further improved its quality by introducing the multi-resolution STFT loss

STFT is calculated in different T/F resolutions

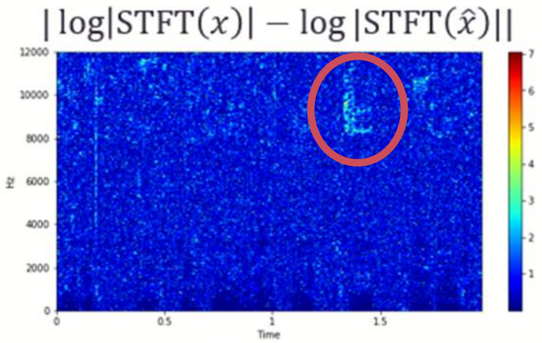
There are two loss functions

One penalizes large energy components

The other penalizes small energy components



Real Spectrogram



Log STFT loss penalizes small amplitude components

$$L_{\mathrm{mr\_stft}}(G) = \frac{1}{M}\sum_{m=1}^{M} L_{\mathrm{stft}}^{(m)}(G)$$

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

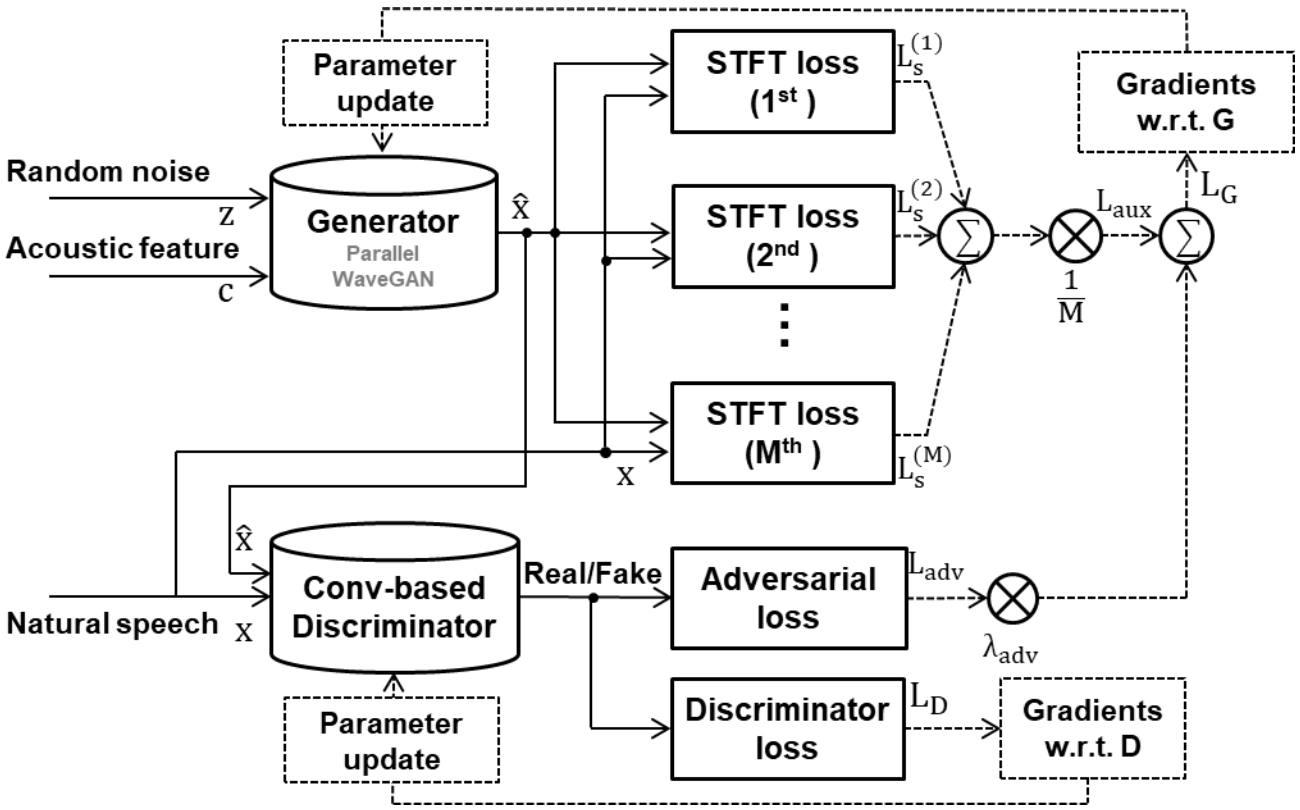
$$L_{\mathrm{sc}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\sqrt{\sum_{t,f}(|\mathbf{X}_{t,f}| - |\hat{\mathbf{X}}_{t,f}|)^2}}{\sqrt{\sum_{t,f}|\mathbf{X}_{t,f}|^2}}$$

$$L_{\mathrm{mag}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\sum_{t,f}|\log|\mathbf{X}_{t,f}| - \log|\hat{\mathbf{X}}_{t,f}||}{T \cdot N}$$


R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

## Training method




R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

## Training method

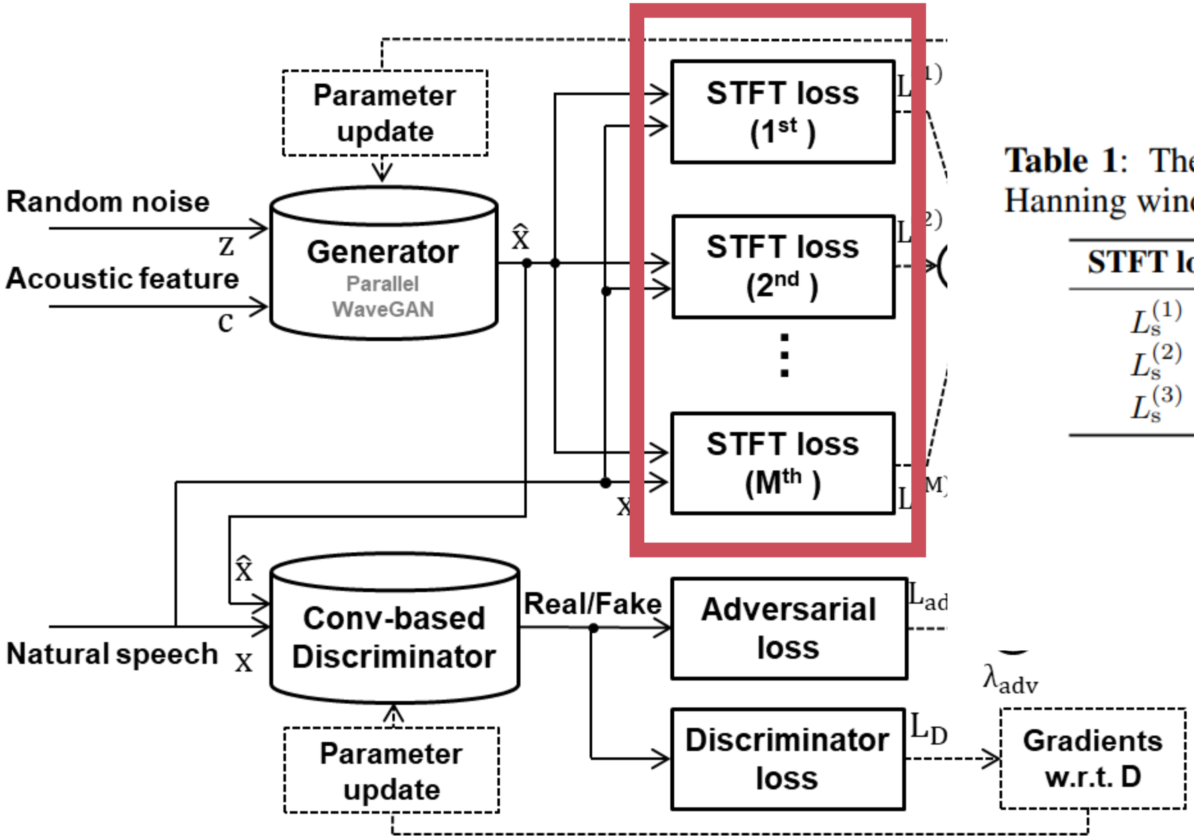



**Table 1**: The details of the multi-resolution STFT loss. A Hanning window was applied before the FFT process.

| STFT loss | FFT size | Window size | Frame shift |
|---|---|---|---|
| $L_s^{(1)}$ | 1024 | 600 (25 ms) | 120 (5 ms) |
| $L_s^{(2)}$ | 2048 | 1200 (50 ms) | 240 (10 ms) |
| $L_s^{(3)}$ | 512 | 240 (10 ms) | 50 (≈ 2 ms) |

R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

## PARALLEL WAVEGAN: A FAST WAVEFORM GENERATION MODEL BASED ON GENERATIVE ADVERSARIAL NETWORKS WITH MULTI-RESOLUTION SPECTROGRAM

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

[1]LINE Corp., Tokyo, Japan.
[2]NAVER Corp., Seongnam, Korea

### ABSTRACT

We propose Parallel WaveGAN, a distillation-free, fast, and small-footprint waveform generation method using a generative adversarial network. In the proposed method, a non-autoregressive WaveNet is trained by jointly optimizing multi-resolution spectrogram and adversarial loss functions, which can effectively capture the time-frequency distribution of the realistic speech waveform. As our method does not require density distillation used in the conventional teacher-student framework, the entire model can be easily trained. Furthermore, our model is able to generate high-fidelity speech even with its compact architecture. In particular, the proposed Parallel WaveGAN has only 1.44 M parameters and can generate 24 kHz speech waveform 28.68 times faster than real-time on a single GPU environment. Perceptual listening test results verify that our proposed method achieves 4.16 mean opinion score within a Transformer-based text-to-speech framework, which is comparative to the best distillation-based Parallel WaveNet system.

# Vocoding models: Parallel WaveGAN

## Evaluation results

**Table 2**: The inference speed and the MOS results with 95% confidence intervals: Acoustic features extracted from the recorded speech signal were used to compose the input auxiliary features. The evaluation was conducted on a server with a single NVIDIA Tesla V100 GPU. Note that the inference speed $k$ means that the system was able to generate waveforms $k$ times faster than real-time.

| System index | Model | KLD-based distillation | STFT loss | Adversarial loss | Number of layers | Model size | Inference speed | MOS |
|---|---|---|---|---|---|---|---|---|
| System 1 | WaveNet | - | - | - | 24 | 3.81 M | $0.32\times10^{-2}$ | $3.61\pm0.12$ |
| System 2 | ClariNet | Yes | $L_s^{(1)}$ | - | 60 | 2.78 M | 14.62 | $3.88\pm0.11$ |
| System 3 | ClariNet | Yes | $L_s^{(1)}+L_s^{(2)}+L_s^{(3)}$ | - | 60 | 2.78 M | 14.62 | $4.21\pm0.09$ |
| System 4 | ClariNet | Yes | $L_s^{(1)}+L_s^{(2)}+L_s^{(3)}$ | Yes | 60 | 2.78 M | 14.62 | $4.21\pm0.09$ |
| System 5 | Parallel WaveGAN | - | $L_s^{(1)}$ | Yes | 30 | 1.44 M | 28.68 | $1.36\pm0.07$ |
| System 6 | Parallel WaveGAN | - | $L_s^{(1)}+L_s^{(2)}+L_s^{(3)}$ | Yes | 30 | 1.44 M | 28.68 | $4.06\pm0.10$ |
| System 7 | Recording | - | - | - | - | - | - | $4.46\pm0.08$ |

**Table 3**: Training time comparison: All the experiments were conducted on a server with two NVIDIA Tesla V100 GPUs. Each vocoder model corresponds to System 1, 3, 4, and 6 described in Table 2, respectively. Note that the times for ClariNets include the training time for the teacher WaveNet.

| Model | Training time (days) |
|---|---|
| WaveNet | 7.4 |
| ClariNet | 12.7 |
| ClariNet-GAN | 13.5 |
| Parallel WaveGAN (ours) | 2.8 |

**Table 4**: MOS results with 95% confidence intervals: Acoustic features generated from the Transformer TTS model were used to compose the input auxiliary features.

| Model | MOS |
|---|---|
| Transformer + WaveNet | $3.33\pm0.11$ |
| Transformer + ClariNet | $4.00\pm0.10$ |
| Transformer + ClariNet-GAN | $4.14\pm0.10$ |
| Transformer + Parallel WaveGAN (ours) | $4.16\pm0.09$ |
| Recording | $4.46\pm0.08$ |

R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

## Evaluation results

**Table 2**: The inference speed and the MOS results with 95% confidence intervals: Acoustic features extracted from the recorded speech signal were used to compose the input auxiliary features. The evaluation was conducted on a server with a single NVIDIA Tesla V100 GPU. Note that the inference speed $k$ means that the system was able to generate waveforms $k$ times faster than real-time.

| System index | Model | KLD-based distillation | STFT loss | Adversarial loss | Number of layers | Model size | Inference speed | MOS |
|---|---|---|---|---|---|---|---|---|
| System 1 | WaveNet | - | - | - | 24 | 3.81 M | $0.32\times10^{-2}$ | $3.61\pm0.12$ |
| System 2 | ClariNet | Yes | $L_s^{(1)}$ | - | 60 | 2.78 M | 14.62 | $3.88\pm0.11$ |
| System 3 | ClariNet | Yes | $L_s^{(1)}+L_s^{(2)}+L_s^{(3)}$ | - | 60 | 2.78 M | 14.62 | $4.21\pm0.09$ |
| System 4 | ClariNet | Yes | $L_s^{(1)}+L_s^{(2)}+L_s^{(3)}$ | Yes | 60 | 2.78 M | 14.62 | $4.21\pm0.09$ |
| System 5 | Parallel WaveGAN | - | $L_s^{(1)}$ | Yes | 30 | 1.44 M | 28.68 | $4.36\pm0.07$ |
| System 6 | Parallel WaveGAN | - | $L_s^{(1)}+L_s^{(2)}+L_s^{(3)}$ | Yes | 30 | 1.44 M | 28.68 | $4.06\pm0.10$ |
| System 7 | Recording | - | - | - | - | | | $4.46\pm0.08$ |

**Table 3**: Training time comparison: All the experiments were conducted on a server with two NVIDIA Tesla V100 GPUs. Each vocoder model corresponds to System 1, 3, 4, and 6 described in Table 2, respectively. Note that the times for ClariNets include the training time for the teacher WaveNet.

| Model | Training time (days) |
|---|---|
| WaveNet | 7.4 |
| ClariNet | 12.7 |
| ClariNet-GAN | 13.5 |
| Parallel WaveGAN (ours) | 2.8 |

**Table 4**: MOS results with 95% confidence intervals: Acoustic features generated from the Transformer TTS model were used to compose the input auxiliary features.

| Model | MOS |
|---|---|
| Transformer + WaveNet | $3.33\pm0.11$ |
| Transformer + ClariNet | $4.00\pm0.10$ |
| Transformer + ClariNet-GAN | $4.14\pm0.10$ |
| Transformer + Parallel WaveGAN (ours) | $4.16\pm0.09$ |
| Recording | $4.46\pm0.08$ |

R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

## Evaluation results

**Table 2**: The inference speed and the MOS results with 95% confidence intervals: Acoustic features extracted from the recorded speech signal were used to compose the input auxiliary features. The evaluation was conducted on a server with a single NVIDIA Tesla V100 GPU. Note that the inference speed $k$ means that the system was able to generate waveforms $k$ times faster than real-time.

| System index | Model | KLD-based distillation | STFT loss | Adversarial loss | Number of layers | Model size | Inference speed | MOS |
|---|---|---|---|---|---|---|---|---|
| System 1 | WaveNet | - | - | - | 24 | 3.81 M | $0.32\times10^{-2}$ | $3.61\pm0.12$ |
| System 2 | ClariNet | Yes | $L_s^{(1)}$ | - | 60 | 2.78 M | 14.62 | $3.88\pm0.11$ |
| System 3 | ClariNet | Yes | $L_s^{(1)}+L_s^{(2)}+L_s^{(3)}$ | - | 60 | 2.78 M | 14.62 | $4.21\pm0.09$ |
| System 4 | ClariNet | Yes | $L_s^{(1)}+L_s^{(2)}+L_s^{(3)}$ | Yes | 60 | 2.78 M | 14.62 | $4.21\pm0.09$ |
| System 5 | Parallel WaveGAN | - | $L_s^{(1)}$ | Yes | 30 | 1.44 M | 28.68 | $1.36\pm0.07$ |
| System 6 | Parallel WaveGAN | - | $L_s^{(1)}+L_s^{(2)}+L_s^{(3)}$ | Yes | 30 | 1.44 M | 28.68 | $4.06\pm0.10$ |
| System 7 | Recording | - | - | - | - | - | - | $4.46\pm0.08$ |

**Table 3**: Training time comparison: All the experiments were conducted on a server with two NVIDIA Tesla V100 GPUs. Each vocoder model corresponds to System 1, 3, 4, and 6 described in Table 2, respectively. Note that the times for ClariNets include the training time for the teacher WaveNet.

| Model | Training time (days) |
|---|---|
| WaveNet | 7.4 |
| ClariNet | 12.7 |
| ClariNet-GAN | 13.5 |
| Parallel WaveGAN (ours) | 2.8 |

**Table 4**: MOS results with 95% confidence intervals: Acoustic features generated from the Transformer TTS model were used to compose the input auxiliary features.

| Model | MOS |
|---|---|
| Transformer + WaveNet | $3.33\pm0.11$ |
| Transformer + ClariNet | $4.00\pm0.10$ |
| Transformer + ClariNet-GAN | $4.14\pm0.10$ |
| Transformer + Parallel WaveGAN (ours) | $4.16\pm0.09$ |
| Recording | $4.46\pm0.08$ |

R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

## Evaluation results

**Table 2**: The inference speed and the MOS results with 95% confidence intervals: Acoustic features extracted from the recorded speech signal were used to compose the input auxiliary features. The evaluation was conducted on a server with a single NVIDIA Tesla V100 GPU. Note that the inference speed $k$ means that the system was able to generate waveforms $k$ times faster than real-time.

| System index | Model | KLD-based distillation | STFT loss | Adversarial loss | Number of layers | Model size | Inference speed | MOS |
|---|---|---|---|---|---|---|---|---|
| System 1 | WaveNet | - | - | - | 24 | 3.81 M | $0.32\times10^{-2}$ | $3.61\pm0.12$ |
| System 2 | ClariNet | Yes | $L_{\mathrm{s}}^{(1)}$ | - | 60 | 2.78 M | 14.62 | $3.88\pm0.11$ |
| System 3 | ClariNet | Yes | $L_{\mathrm{s}}^{(1)} + L_{\mathrm{s}}^{(2)} + L_{\mathrm{s}}^{(3)}$ | - | 60 | 2.78 M | 14.62 | $4.21\pm0.09$ |
| System 4 | ClariNet | Yes | $L_{\mathrm{s}}^{(1)} + L_{\mathrm{s}}^{(2)} + L_{\mathrm{s}}^{(3)}$ | Yes | 60 | 2.78 M | 14.62 | $4.21\pm0.09$ |
| System 5 | Parallel WaveGAN | - | $L_{\mathrm{s}}^{(1)}$ | Yes | 30 | 1.44 M | 28.68 | $1.36\pm0.07$ |
| System 6 | Parallel WaveGAN | - | $L_{\mathrm{s}}^{(1)} + L_{\mathrm{s}}^{(2)} + L_{\mathrm{s}}^{(3)}$ | Yes | 30 | 1.44 M | 28.68 | $4.06\pm0.10$ |
| System 7 | Recording | - | - | - | - | - | - | $4.46\pm0.08$ |

**Table 3**: Training time comparison: All the experiments were conducted on a server with two NVIDIA Tesla V100 GPUs. Each vocoder model corresponds to System 1, 3, 4, and 6 described in Table 2, respectively. Note that the times for ClariNets include the training time for the teacher WaveNet.

| Model | Training time (days) |
|---|---|
| WaveNet | 7.4 |
| ClariNet | 12.7 |
| ClariNet-GAN | 13.5 |
| Parallel WaveGAN (ours) | 2.8 |

**Table 4**: MOS results with 95% confidence intervals: Acoustic features generated from the Transformer TTS model were used to compose the input auxiliary features.

| Model | MOS |
|---|---|
| Transformer + WaveNet | $3.33\pm0.11$ |
| Transformer + ClariNet | $4.00\pm0.10$ |
| Transformer + ClariNet-GAN | $4.14\pm0.10$ |
| Transformer + Parallel WaveGAN (ours) | $4.16\pm0.09$ |
| Recording | $4.46\pm0.08$ |

R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

# Vocoding models: Parallel WaveGAN

Demo samples

Open source
(implemented by Tomoki Hayashi, Nagoya Univ.)

R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6194-6198.

# Parallel waveform synthesis

Parallel WaveGAN: Toward high-quality synthesis

# Toward high-quality synthesis

## IMPROVED PARALLEL WAVEGAN VOCODER WITH PERCEPTUALLY WEIGHTED SPECTROGRAM LOSS

*Eunwoo Song*[1], *Ryuichi Yamamoto*[2], *Min-Jae Hwang*[3], *Jin-Seob Kim*[1], *Ohsung Kwon*[1], *Jae-Min Kim*[1]

[1]NAVER Corp., Seongnam, Korea
[2]LINE Corp., Tokyo, Japan
[3]Search Solutions Inc., Seongnam, Korea

**ABSTRACT**

This paper proposes a spectral-domain perceptual weighting technique for Parallel WaveGAN-based text-to-speech (TTS) systems. The recently proposed Parallel WaveGAN vocoder successfully generates waveform sequences using a fast non-autoregressive WaveNet model. By employing multi-resolution short-time Fourier transform (MR-STFT) criteria with a generative adversarial network, the light-weight convolutional networks can be effectively trained without any distillation process. To further improve the vocoding performance, we propose the application of frequency-dependent weighting to the MR-STFT loss function. The proposed method penalizes perceptually-sensitive errors in the frequency domain; thus, the model is optimized toward reducing auditory noise in the synthesized speech. Subjective listening test results demonstrate that our proposed method achieves 4.21 and 4.26 TTS mean opinion scores for female and male Korean speakers, respectively.

"Weighted spectral Loss"

# Toward high-quality synthesis

## PARALLEL WAVEFORM SYNTHESIS BASED ON GENERATIVE ADVERSARIAL NETWORKS WITH VOICING-AWARE CONDITIONAL DISCRIMINATORS

*Ryuichi Yamamoto*[1], *Eunwoo Song*[2], *Min-Jae Hwang*[3] *and Jae-Min Kim*[2]

[1]LINE Corp., Tokyo, Japan
[2]NAVER Corp., Seongnam, Korea
[3]Search Solutions Inc., Seongnam, Korea

**ABSTRACT**

This paper proposes voicing-aware conditional discriminators for Parallel WaveGAN-based waveform synthesis systems. In this framework, we adopt a projection-based conditioning method that can significantly improve the discriminator's performance. Furthermore, the conventional discriminator is separated into two waveform discriminators for modeling voiced and unvoiced speech. As each discriminator learns the distinctive characteristics of the harmonic and noise components, respectively, the adversarial training process becomes more efficient, allowing the generator to produce more realistic speech waveforms. Subjective test results demonstrate the superiority of the proposed method over the conventional Parallel WaveGAN and WaveNet systems. In particular, our speaker-independently trained model within a FastSpeech 2 based text-to-speech framework achieves the mean opinion scores of 4.20, 4.18, 4.21, and 4.31 for four Japanese speakers, respectively.

"Voicing-aware discriminators"

# Toward high-quality synthesis

## High-fidelity Parallel WaveGAN with Multi-band Harmonic-plus-Noise Model

*Min-Jae Hwang*[1]*, *Ryuichi Yamamoto*[2]*, *Eunwoo Song*[3] and *Jae-Min Kim*[3]

[1]Search Solutions Inc., Seongnam, Korea
[2]LINE Corp.,Tokyo, Japan
[3]NAVER Corp., Seongnam, Korea

### Abstract

This paper proposes a multi-band harmonic-plus-noise (HN) Parallel WaveGAN (PWG) vocoder. To generate a high-fidelity speech signal, it is important to well-reflect the harmonic-noise characteristics of the speech waveform in the time-frequency domain. However, it is difficult for the conventional PWG model to accurately match this condition, as its single generator inefficiently represents the complicated nature of harmonic-noise structures. In the proposed method, the HN WaveNet models are employed to overcome this limitation, which enable the separate generation of the harmonic and noise components of speech signals from the pitch-dependent sine wave and Gaussian noise sources, respectively. Then, the energy ratios between harmonic and noise components in multiple frequency bands (i.e., subband harmonicities) are predicted by an additional harmonicity estimator. Weighted by the estimated harmonicities, the gain of harmonic and noise components in each subband is adjusted, and finally mixed together to compose the full-band speech signal. Subjective evaluation results showed that the proposed method significantly improved the perceptual quality of the synthesized speech.

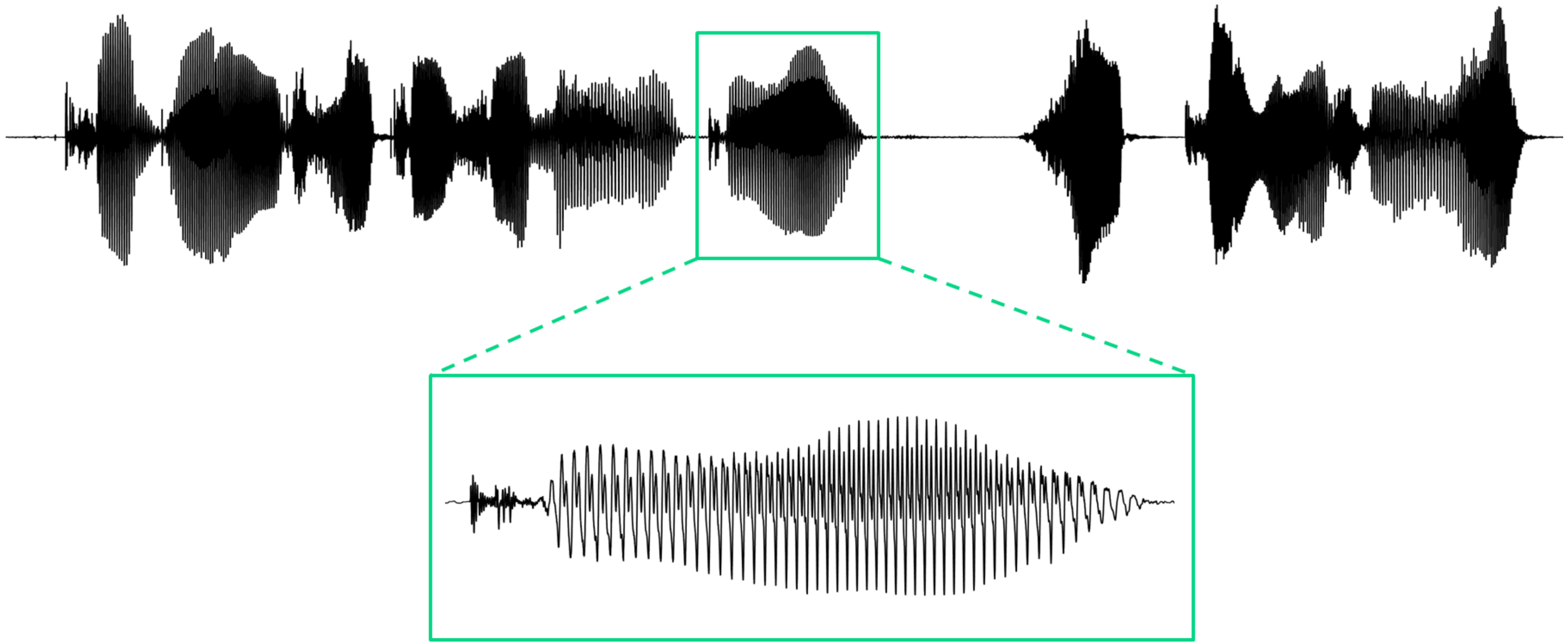
"Harmonic/noise generators"

# Parallel waveform synthesis

Toward high-quality synthesis: Speech fundamentals

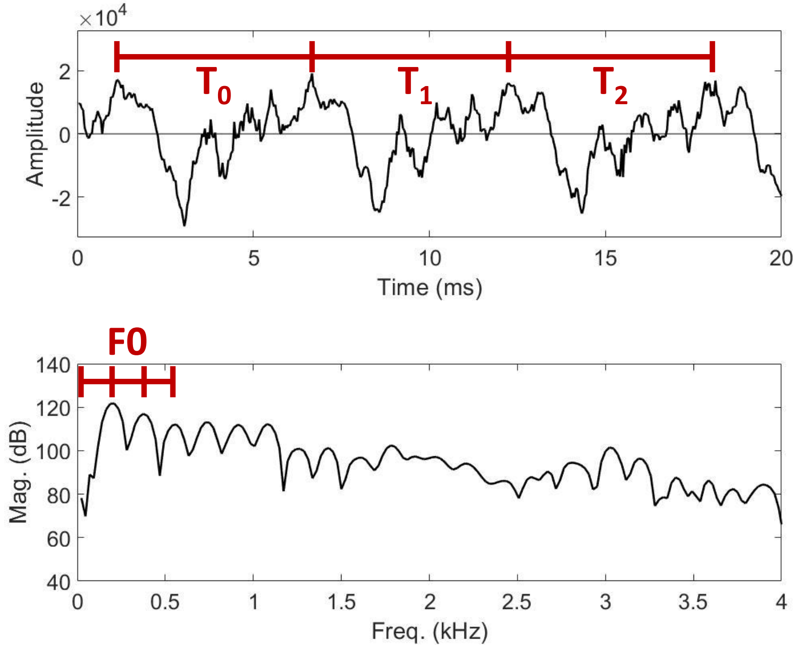
# Speech fundamentals

## Speech waveform



# Speech fundamentals

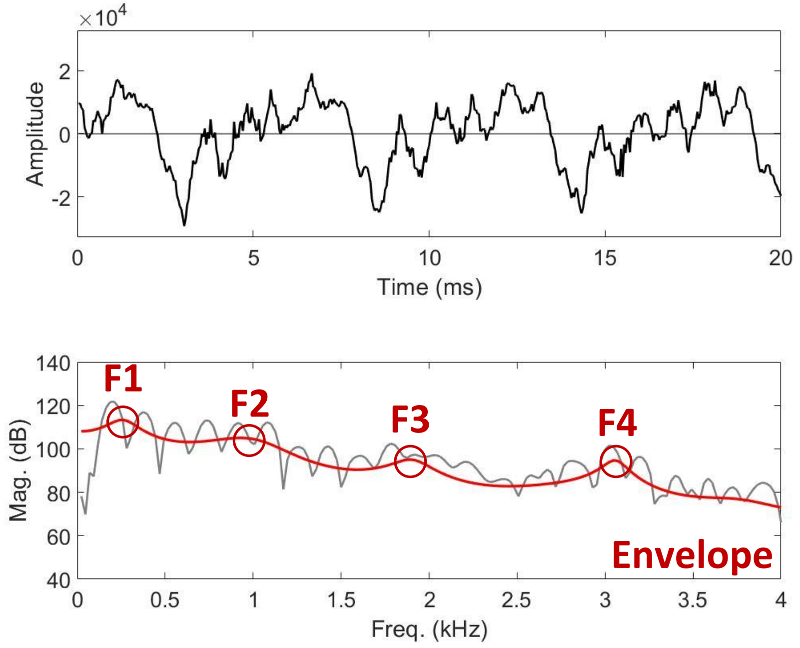
## Pitch period




- Pitch period = $T\_0 \approx T\_1 \approx T\_2$
  - Long-term period of speech (time-domain)
- Fundamental frequency (F0) = $1/T\_0$
  - 1 / PP (frequency-domain)
  - Female voice : Ave. 200 Hz
  - Male voice : Ave. 100 Hz
- Harmonic spectrum
  - Multiple peaks of speech spectrum (interval=F0)
- Formant frequency (F1, F2, …)
  - Vocal tract resonance

# Speech fundamentals

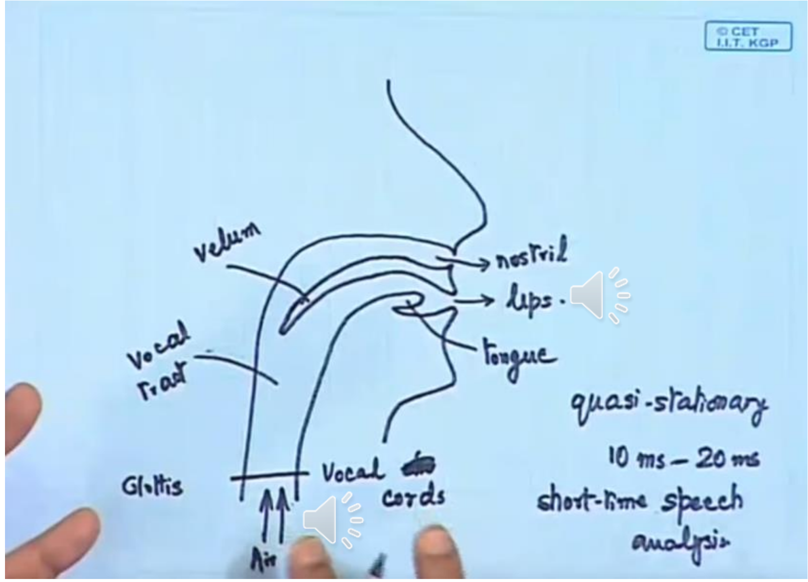
## Formant frequency




- Pitch period = $T\_0 \approx T\_1 \approx T\_2$
  - Long-term period of speech (time-domain)
- Fundamental frequency (F0) = $1/T\_0$
  - 1 / PP (frequency-domain)
  - Female voice : Ave. 200 Hz
  - Male voice : Ave. 100 Hz
- Harmonic spectrum
  - Multiple peaks of speech spectrum (interval=F0)
- Formant frequency (F1, F2, …)
  - Vocal tract resonance

# How do we produce speech?

## Speech production model



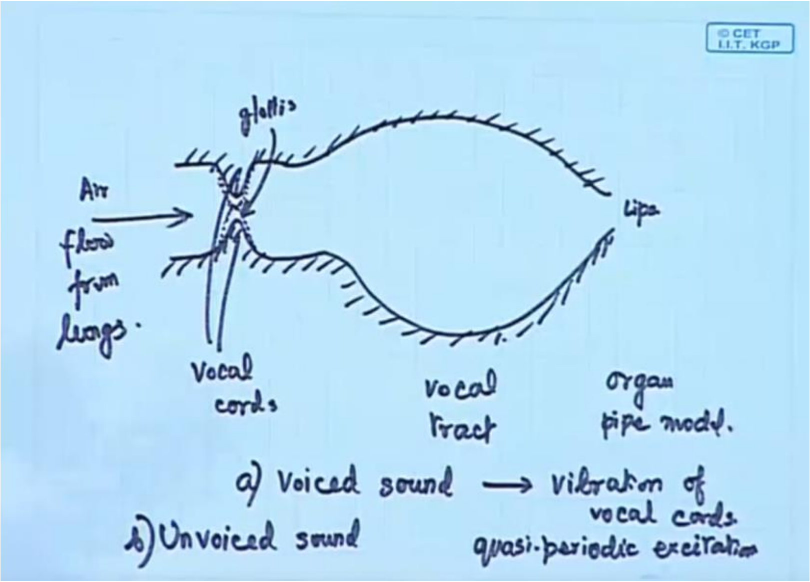

https://www.youtube.com/watch?v=X_JvfZiGEek

- Lung
  - Power supply
- Vocal source
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract filter
  - Shaping voice color

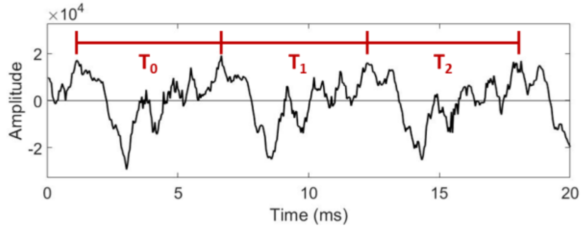
**Source** → **Filter** → **Speech**

# How do we produce speech?

## Speech production model




https://www.youtube.com/watch?v=X_JvfZiGEek

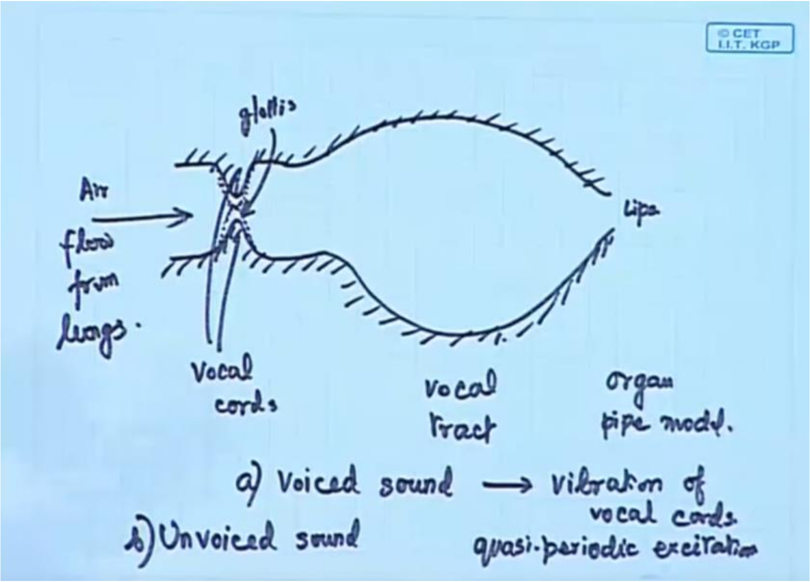



- Lung
  - Power supply
- Vocal source
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract filter
  - Shaping voice color

Source → **Filter** → Speech

# How do we produce speech?

## Speech production model



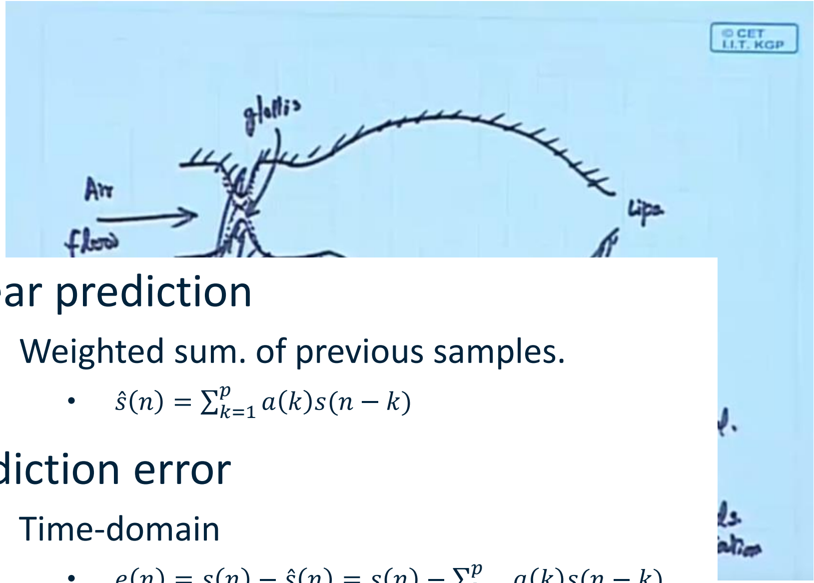

https://www.youtube.com/watch?v=X_JvfZiGEek

- Lung
  - Power supply
- Vocal source
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract filter
  - Shaping voice color

Source → $\boxed{\text{Filter}}$ → Speech

# How do we produce speech?

## Speech production model



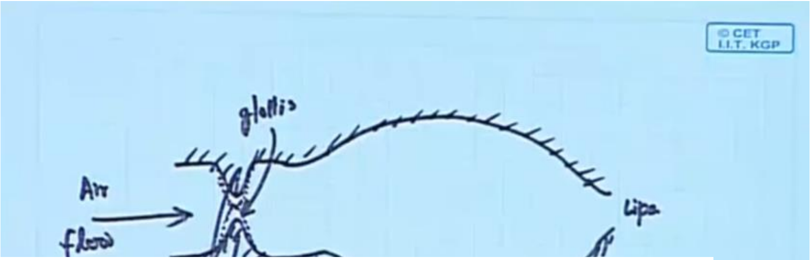

- Linear prediction
  - Weighted sum. of previous samples.
    - $\hat{s}(n) = \sum_{k=1}^{p} a(k)s(n-k)$
- Prediction error
  - Time-domain
    - $e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a(k)s(n-k)$
  - Minimizing mean square error
    - $\underset{a_k}{\operatorname{argmin}} E\left\{\left\|s(n) - \sum_{k=1}^{p} a(k)s(n-k)\right\|^2\right\}$

- Lung
  - Power supply
- Vocal source
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract filter
  - Shaping voice color

Source → $\boxed{\text{Filter}}$ → Speech

# How do we produce speech?

## Speech production model




- Linear prediction
  - Weighted sum. of previous samples.
    - $\hat{s}(n) = \sum_{k=1}^{p} a(k)s(n-k)$
- Prediction error
  - Frequency-domain
    - $S(z) = E(z)H(z) = E(z)\boxed{\frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}}}$

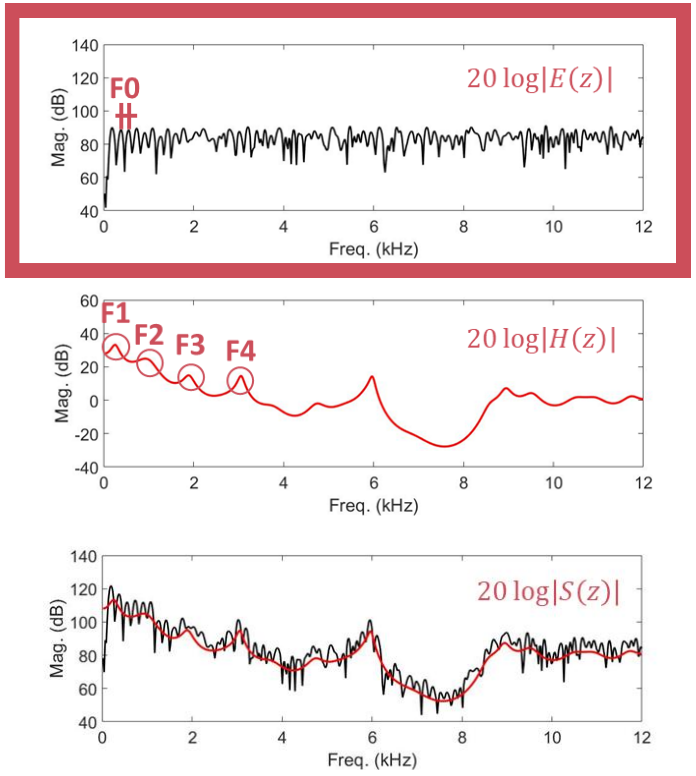
LPC filter

- Lung
  - Power supply
- Vocal source
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract filter
  - Shaping voice color

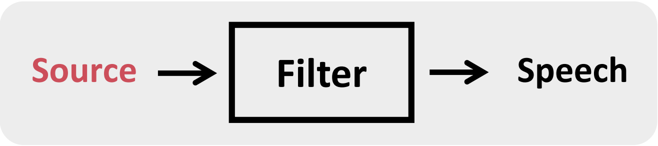
Source → Filter → Speech

# How do we produce speech?

## Speech production model

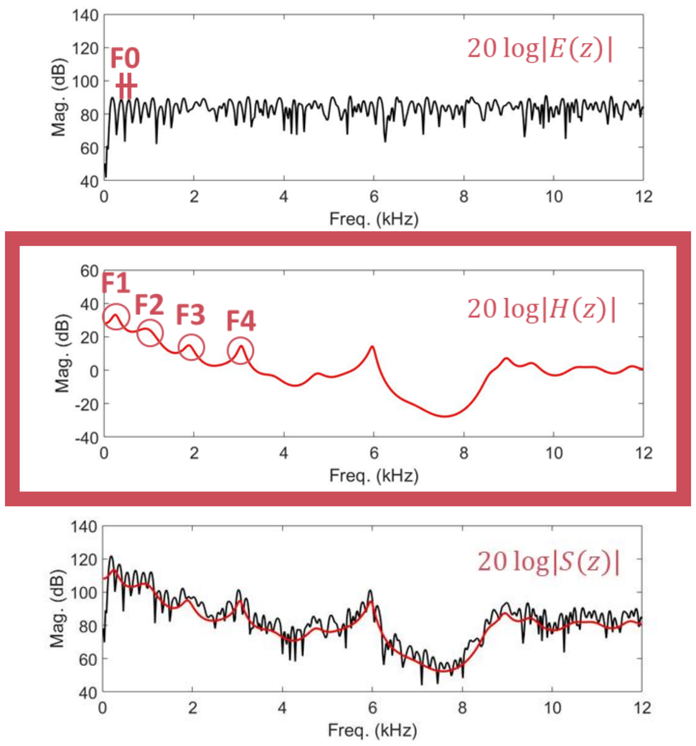



- Lung
  - Power supply
- Vocal source
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract filter
  - Shaping voice color



# How do we produce speech?

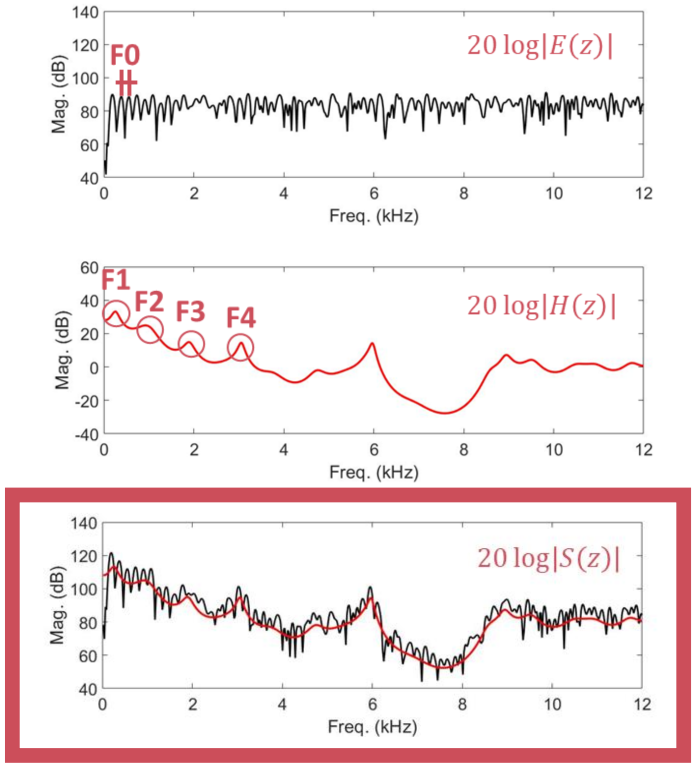
## Speech production model




- Lung
  - Power supply
- Vocal source
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract filter
  - Shaping voice color

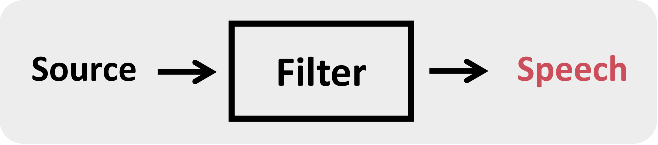
# How do we produce speech?

## Speech production model




$$S(z) = E(z)H(z) = E(z) \times \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

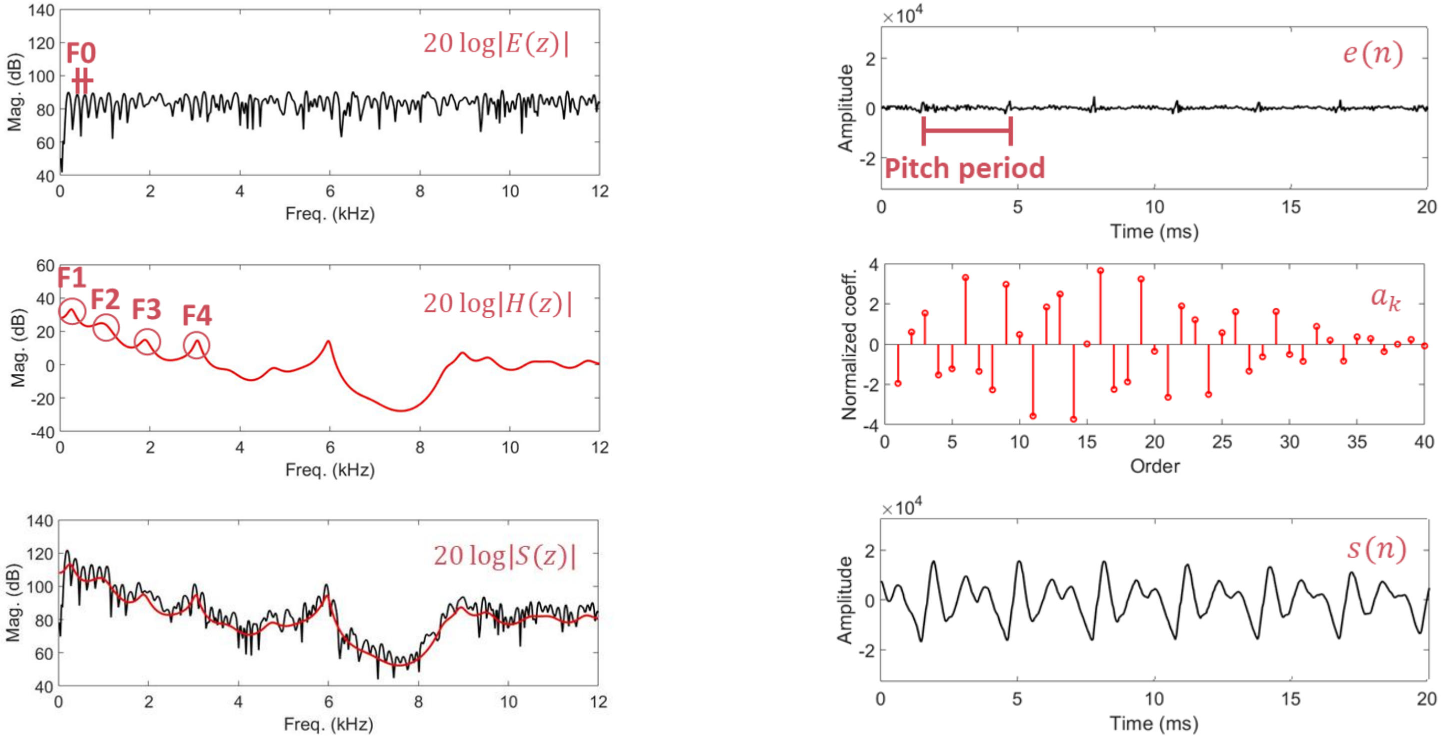
- Lung
  - Power supply
- Vocal source
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract filter
  - Shaping voice color



# How do we produce speech?

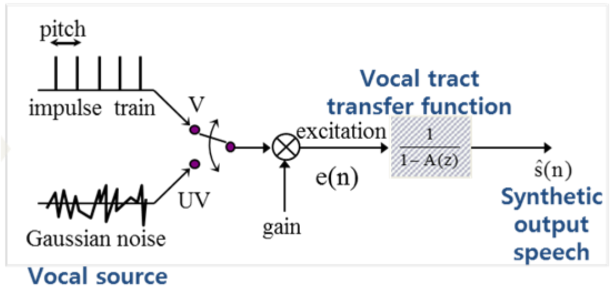
## Speech production model

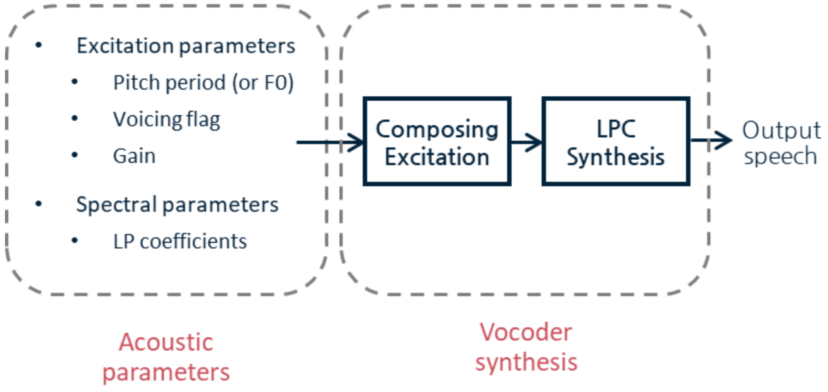
→ Time-domain

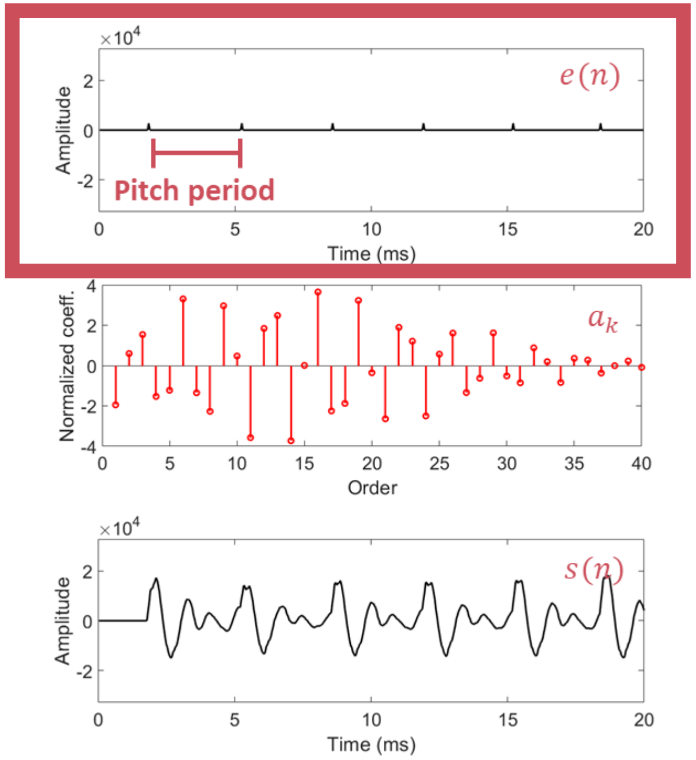


# How do we produce speech?

## Parametric LPC vocoder



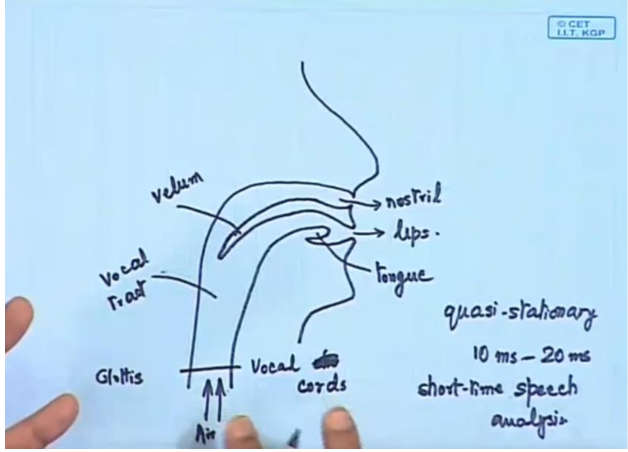



# Parallel waveform synthesis

Toward high-quality synthesis: Perceptually weighted spectral loss

# Perceptually weighted spectral loss

## Combining LPC synthesis filter with neural excitation vocoders


https://www.youtube.com/watch?v=X_JvfZiGEek

### Speech production model

Vocal source → Excitation

Voiced sound: quasi-periodic
Unvoiced sound: aperiodic

Vocal tract → LPC synthesis

Shaping voice color

E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” Proc. SLT, 2021, pp. 470-476.

# Perceptually weighted spectral loss

**Combining LPC synthesis filter with neural excitation vocoders**


WaveNet + LPC filter = ExcitNet, LP-WaveNet, …

WaveRNN + LPC filter = LPCNet

E. Song, et al., "Improved Parallel WaveGAN with perceptually weighted spectrogram loss," Proc. SLT, 2021, pp. 470-476.

# Perceptually weighted spectral loss

**Combining LPC synthesis filter with neural excitation vocoders**


WaveNet + LPC filter = ExcitNet, LP-WaveNet, …

WaveRNN + LPC filter = LPCNet

WaveGlow + LPC filter = ?

Parallel WaveGAN + LPC filter = ?

E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” Proc. SLT, 2021, pp. 470-476.

# Perceptually weighted spectral loss

## Combining LPC synthesis filter with neural excitation vocoders


Autoregressive models

WaveNet + LPC filter = ExcitNet, LP-WaveNet, …

WaveRNN + LPC filter = LPCNet

WaveGlow + LPC filter = ?

Parallel WaveGAN + LPC filter = ?

Non-autoregressive models

E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” Proc. SLT, 2021, pp. 470-476.

# Perceptually weighted spectral loss

**Combining LPC synthesis filter with neural excitation vocoders**


Autoregressive models

WaveNet + LPC filter = ExcitNet, LP-WaveNet, …

WaveRNN + LPC filter = LPCNet

WaveGlow + LPC filter = ?

Parallel WaveGAN + LPC filter = ?

Non-autoregressive models

→ Not suitable for estimating excitation signals

E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” Proc. SLT, 2021, pp. 470-476.

# Recall: Parallel WaveGAN

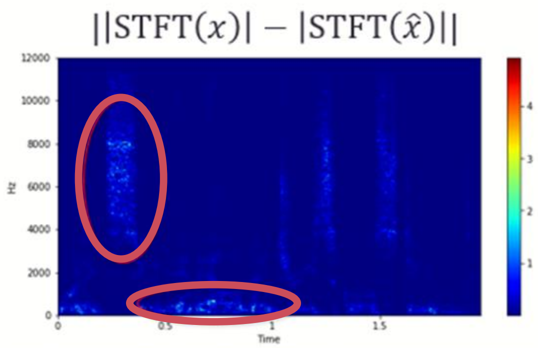
1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method
3. Further improved its quality by introducing the multi-resolution STFT loss

STFT is calculated in different T/F resolutions

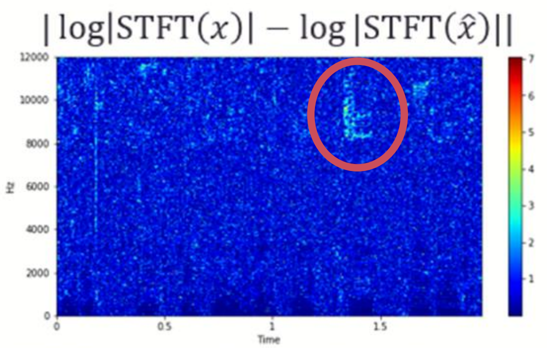
There are two loss functions

One penalizes large energy components

The other penalizes small energy components



SC penalizes large amplitude components



Log STFT loss penalizes small amplitude components

$$L_{\mathrm{mr\_stft}}(G) = \frac{1}{M}\sum_{m=1}^{M} L_{\mathrm{stft}}^{(m)}(G)$$

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

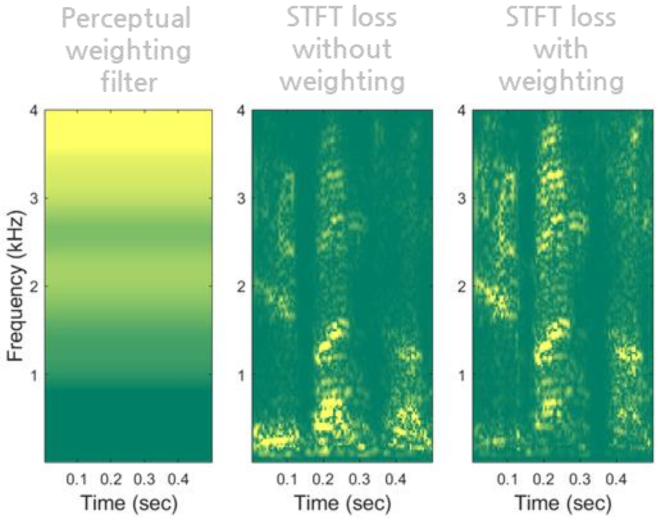
$$L_{\mathrm{sc}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\sqrt{\sum_{t,f}(|\mathbf{X}_{t,f}| - |\hat{\mathbf{X}}_{t,f}|)^2}}{\sqrt{\sum_{t,f}|\mathbf{X}_{t,f}|^2}}$$

$$L_{\mathrm{mag}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\sum_{t,f}|\log|\mathbf{X}_{t,f}| - \log|\hat{\mathbf{X}}_{t,f}||}{T \cdot N}$$


E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” Proc. SLT, 2021, pp. 470-476.

# Perceptually weighted spectral loss

1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method
3. Further improved its quality by introducing the multi-resolution STFT loss
   + Applying perceptual weighting filter




$$L^{w}_{\mathrm{sc}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\sqrt{\sum_{t,f}(\mathbf{W}_{t,f}(|\mathbf{X}_{t,f}| - |\hat{\mathbf{X}}_{t,f}|))^2}}{\sqrt{\sum_{t,f}|\mathbf{X}_{t,f}|^2}}$$

$$L^{w}_{\mathrm{mag}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\sum_{t,f}|\log \mathbf{W}_{t,f}(\log|\mathbf{X}_{t,f}| - \log|\hat{\mathbf{X}}_{t,f}|)|}{T \cdot N}$$

$$W(z) = 1 - \sum_{k=1}^{p} \tilde{\alpha}_k z^{-k}$$

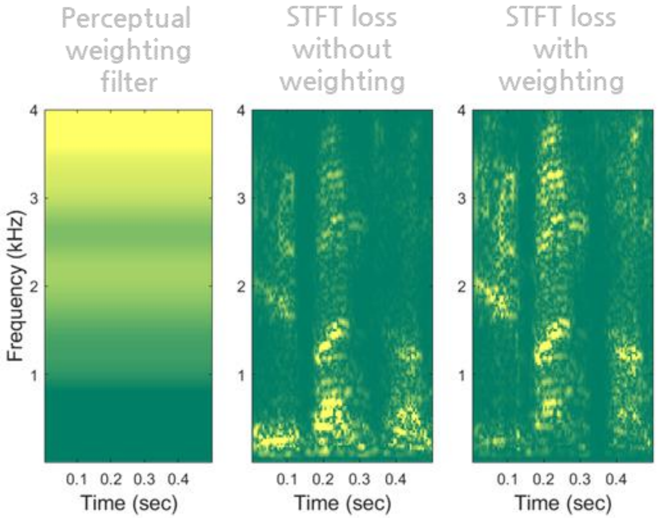

E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” Proc. SLT, 2021, pp. 470-476.

# Perceptually weighted spectral loss

1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method
3. Further improved its quality by introducing the multi-resolution STFT loss

+ Applying perceptual weighting filter

This penalizes perceptually-sensitive errors in the freq. domain




$$L_{\mathrm{sc}}^{w}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sqrt{\sum_{t,f}(\mathbf{W}_{t,f}(|\mathbf{X}_{t,f}| - |\hat{\mathbf{X}}_{t,f}|))^2}}{\sqrt{\sum_{t,f}|\mathbf{X}_{t,f}|^2}}$$

$$L_{\mathrm{mag}}^{w}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sum_{t,f}|\log \mathbf{W}_{t,f}(\log|\mathbf{X}_{t,f}| - \log|\hat{\mathbf{X}}_{t,f}|)|}{T \cdot N}$$

$$W(z) = 1 - \sum_{k=1}^{p} \tilde{\alpha}_k z^{-k}$$

E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” Proc. SLT, 2021, pp. 470-476.

# Perceptually weighted spectral loss

## Evaluation results

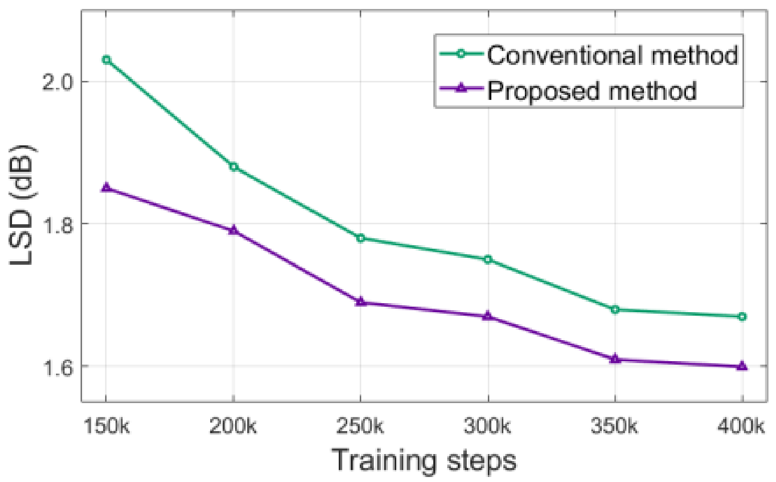



**Fig. 2**: Log-spectral distance (LSD; dB) between the original and generated speech signals

**Table 4**: Naturalness MOS test results with 95% confidence intervals for the TTS systems with respect to the different vocoding models: The MOS results for the proposed system are in bold font. The KRF and KRM denote Korean female and male speakers, respectively.

| Index | Model | KRF | KRM |
|---|---|---|---|
| Test 1 | WaveNet | 3.64±0.14 | 3.60±0.13 |
| Test 2 | WaveNet + NS | 4.36±0.11 | 4.32±0.10 |
| Test 3 | Parallel WaveGAN | 4.02±0.10 | 4.11±0.11 |
| Test 4 | Parallel WaveGAN + NS | 2.34±0.10 | 1.72±0.09 |
| **Test 5** | **Parallel WaveGAN + PW** | **4.26±0.10** | **4.21±0.10** |
| Test 6 | Raw | 4.64±0.07 | 4.59±0.09 |

Acoustic model: Tacotron 2
NS: Noise-shaping (similar to LPC synthesis)

E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” Proc. SLT, 2021, pp. 470-476.

# Perceptually weighted spectral loss

## Evaluation results

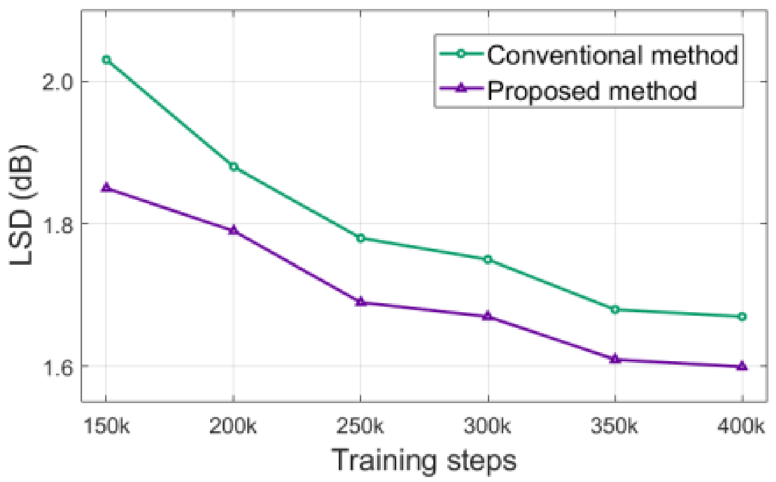



**Fig. 2**: Log-spectral distance (LSD; dB) between the original and generated speech signals

**Table 4**: Naturalness MOS test results with 95% confidence intervals for the TTS systems with respect to the different vocoding models: The MOS results for the proposed system are in bold font. The KRF and KRM denote Korean female and male speakers, respectively.

| Index | Model | KRF | KRM |
|---|---|---|---|
| Test 1 | WaveNet | $3.64 \pm 0.14$ | $3.60 \pm 0.13$ |
| Test 2 | WaveNet + NS | $4.36 \pm 0.11$ | $4.32 \pm 0.10$ |
| Test 3 | Parallel WaveGAN | $4.02 \pm 0.10$ | $4.11 \pm 0.11$ |
| Test 4 | Parallel WaveGAN + NS | $2.34 \pm 0.10$ | $1.72 \pm 0.09$ |
| **Test 5** | **Parallel WaveGAN + PW** | $\mathbf{4.26 \pm 0.10}$ | $\mathbf{4.21 \pm 0.10}$ |
| Test 6 | Raw | $4.64 \pm 0.07$ | $4.59 \pm 0.09$ |

Acoustic model: Tacotron 2
NS: Noise-shaping (similar to LPC synthesis)


E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” Proc. SLT, 2021, pp. 470-476.

# Perceptually weighted spectral loss

## Evaluation results

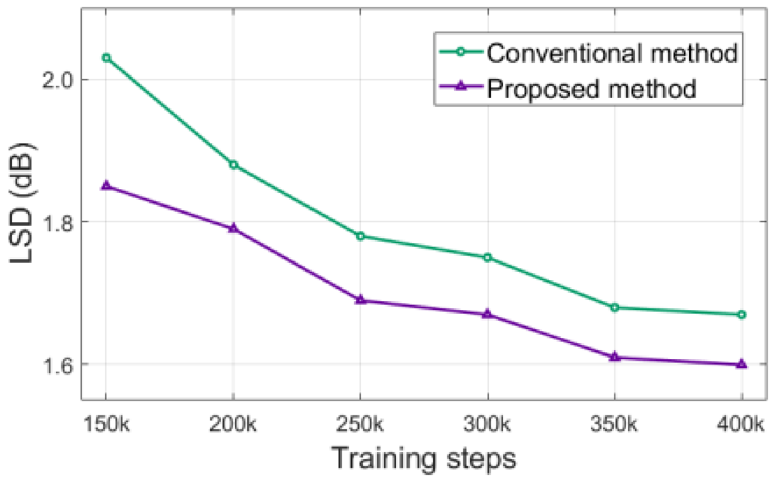



**Fig. 2**: Log-spectral distance (LSD; dB) between the original and generated speech signals

**Table 4**: Naturalness MOS test results with 95% confidence intervals for the TTS systems with respect to the different vocoding models: The MOS results for the proposed system are in bold font. The KRF and KRM denote Korean female and male speakers, respectively.

| Index | Model | KRF | KRM |
|---|---|---|---|
| Test 1 | WaveNet | $3.64 \pm 0.14$ | $3.60 \pm 0.13$ |
| Test 2 | WaveNet + NS | $4.36 \pm 0.11$ | $4.32 \pm 0.10$ |
| Test 3 | Parallel WaveGAN | $4.02 \pm 0.10$ | $4.11 \pm 0.11$ |
| Test 4 | Parallel WaveGAN + NS | $2.34 \pm 0.10$ | $1.72 \pm 0.09$ |
| **Test 5** | **Parallel WaveGAN + PW** | $\mathbf{4.26 \pm 0.10}$ | $\mathbf{4.21 \pm 0.10}$ |
| Test 6 | Raw | $4.64 \pm 0.07$ | $4.59 \pm 0.09$ |

Acoustic model: Tacotron 2
NS: Noise-shaping (similar to LPC synthesis)

E. Song, et al., "Improved Parallel WaveGAN with perceptually weighted spectrogram loss," Proc. SLT, 2021, pp. 470-476.

# Perceptually weighted spectral loss

## Evaluation results

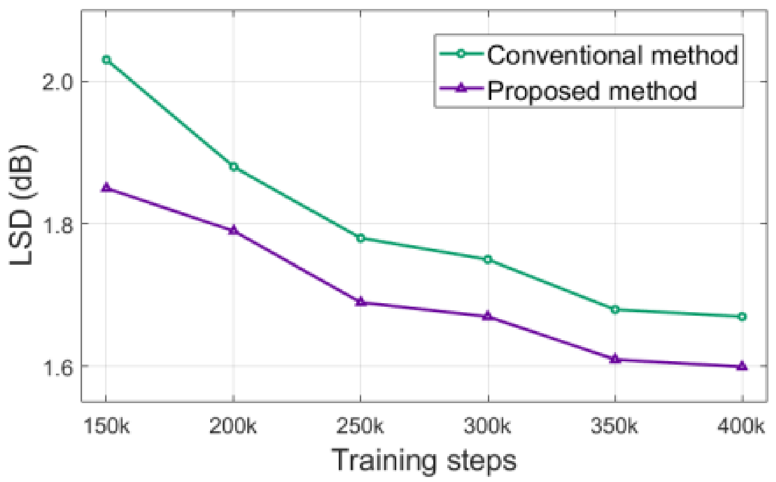



**Fig. 2**: Log-spectral distance (LSD; dB) between the original and generated speech signals

**Table 4**: Naturalness MOS test results with 95% confidence intervals for the TTS systems with respect to the different vocoding models: The MOS results for the proposed system are in bold font. The KRF and KRM denote Korean female and male speakers, respectively.

| Index | Model | KRF | KRM |
|---|---|---|---|
| Test 1 | WaveNet | 3.64±0.14 | 3.60±0.13 |
| Test 2 | WaveNet + NS | 4.36±0.11 | 4.32±0.10 |
| Test 3 | Parallel WaveGAN | 4.02±0.10 | 4.11±0.11 |
| Test 4 | Parallel WaveGAN + NS | 2.34±0.10 | 1.72±0.09 |
| **Test 5** | **Parallel WaveGAN + PW** | **4.26±0.10** | **4.21±0.10** |
| Test 6 | Raw | 4.64±0.07 | 4.59±0.09 |

Acoustic model: Tacotron 2
NS: Noise-shaping (similar to LPC synthesis)

E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” Proc. SLT, 2021, pp. 470-476.

# Perceptually weighted spectral loss

Demo samples

E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” Proc. SLT, 2021, pp. 470-476.

# Parallel waveform synthesis

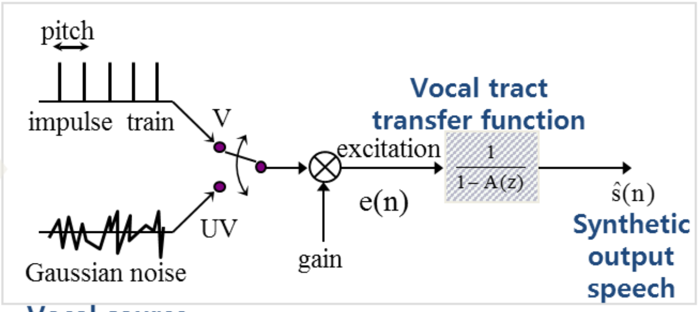
Toward high-quality synthesis: Voicing-aware discriminators

# Voicing-aware discriminators

## Voiced/unvoiced sounds

Voiced sound: Quasi-periodic

Unvoiced sound: aperiodic




R. Yamamoto, et al., “Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators,” Proc. ICASSP, 2021, pp. 6039-6043.

# Voicing-aware discriminators

## Voiced/unvoiced sounds

Voiced sound: Quasi-periodic

Unvoiced sound: aperiodic

V: Characterized by slowly evolving harmonic components

Discriminator should cover long-term variations of voiced sound

R. Yamamoto, et al., “Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators,” Proc. ICASSP, 2021, pp. 6039-6043.

# Voicing-aware discriminators

**Voiced/unvoiced sounds**

Voiced sound: Quasi-periodic

Unvoiced sound: aperiodic

V: Characterized by slowly evolving harmonic components

Discriminator should cover long-term variations of voiced sound

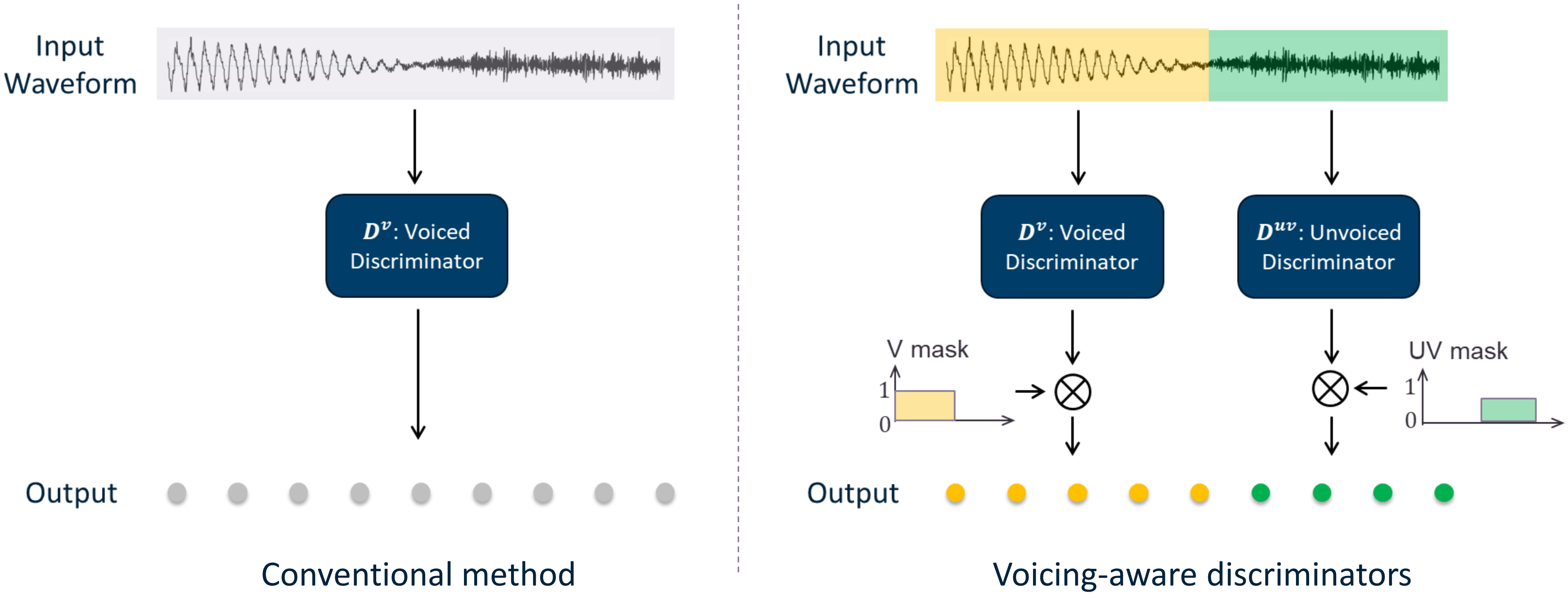
UV: Characterized by rapidly evolving noise components

Discriminator should catch short-term variations of unvoiced sound

R. Yamamoto, et al., “Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators,” Proc. ICASSP, 2021, pp. 6039-6043.

# Voicing-aware discriminators

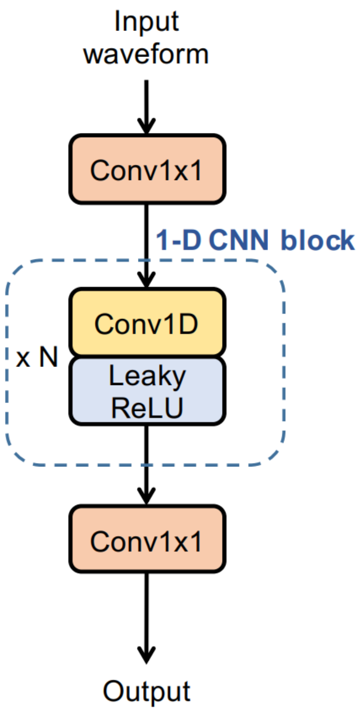
## Voiced/unvoiced masking



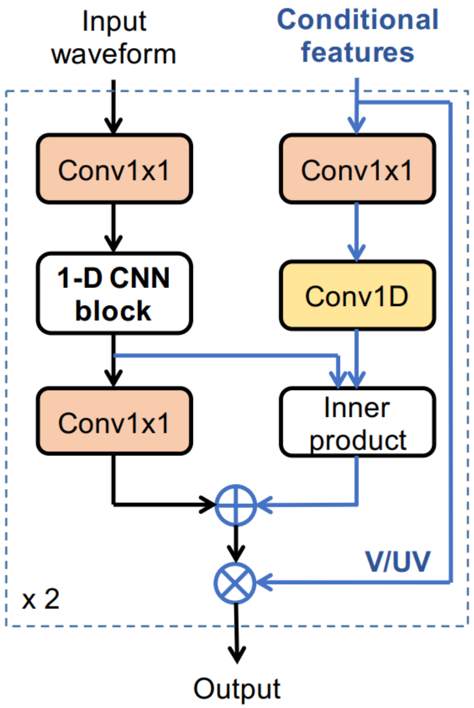

R. Yamamoto, et al., "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," Proc. ICASSP, 2021, pp. 6039-6043.

# Voicing-aware discriminators

## Voiced/unvoiced masking




Conventional method




Voicing-aware discriminators

T. Miyato, et al., “cGANs with projection discriminator,” Proc. ICLR, 2018.

R. Yamamoto, et al., “Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators,” Proc. ICASSP, 2021, pp. 6039-6043.

# Voicing-aware discriminators

## Receptive field

**Table 1.** The dilation factors and receptive fields in the 1-D CNN blocks of the voicing-aware discriminators.

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

Voiced discriminator

- Dilated convolution with long receptive field
- Covering long-term variations of voiced sound

Unvoiced discriminator

- Non-dilated convolution with short receptive field
- Catching short-term variations of unvoiced sound

R. Yamamoto, et al., “Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators,” Proc. ICASSP, 2021, pp. 6039-6043.

# Voicing-aware discriminators

## Receptive field

**Table 1**. The dilation f
blocks of the voicing-aw

| **Discriminator** |
|---|
| $D^{\mathrm{v}}$ |
| $D^{\mathrm{uv}}$ |

Voiced discriminator

- Dilated convoluti
- Covering long-ter

Unvoiced discriminato

- Non-dilated conv
- Catching short-te

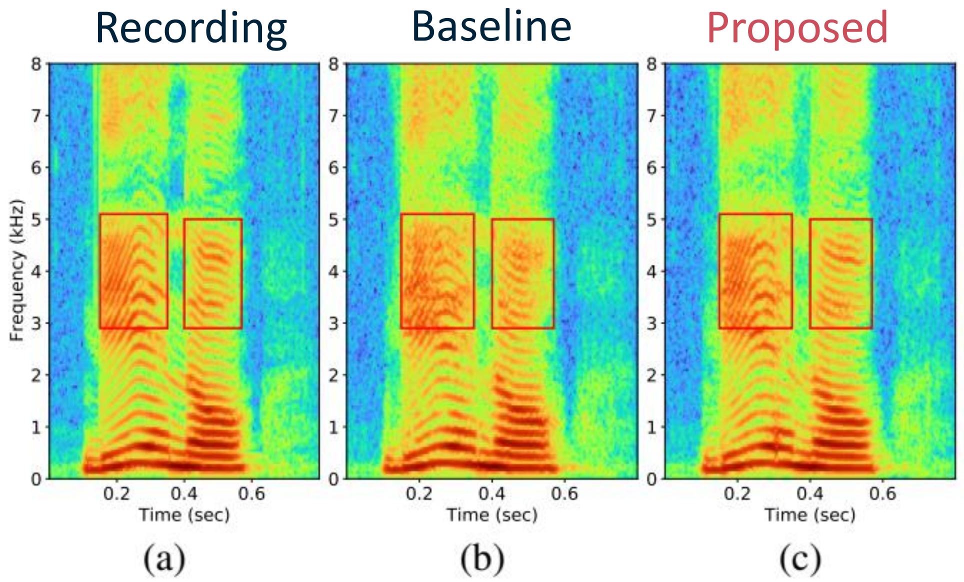



**Fig. 2**. Spectrograms of (a) natural speech, (b) generated speech from the conventional Parallel WaveGAN (S2), and (c) generated speech from the proposed Parallel WaveGAN (S7). As demonstrated in rectangle areas, our proposed method is able to model spectral harmonics more accurately.

R. Yamamoto, et al., “Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators,” Proc. ICASSP, 2021, pp. 6039-6043.

# Voicing-aware discriminators

## Evaluation results

**Table 2.** MOS test results with 95% confidence intervals in analysis/synthesis: The speech samples were generated using the acoustic features extracted from the recorded speech. PWG denotes Parallel WaveGAN for short. Note that systems S2 and S3 used $D^{\mathrm{v}}$ as the primary discriminator. All the models were trained in a speaker-independent manner.

| System | Model | Voiced segments | Unvoiced segments | Discriminator conditioning | MOS F1 | MOS F2 | MOS M1 | MOS M2 |
|---|---|---|---|---|---|---|---|---|
| S1 | WaveNet | - | - | - | 3.64±0.12 | 3.83±0.11 | 3.33±0.12 | 3.13±0.11 |
| S2 | PWG | - | - | - | 3.61±0.11 | 3.55±0.11 | 3.57±0.12 | 3.61±0.11 |
| S3 | PWG-cGAN-D | - | - | Yes | 4.04±0.10 | 3.95±0.10 | 3.91±0.11 | 3.97±0.10 |
| S4 | PWG-V/UV-D | $D^{\mathrm{v}}$ | $D^{\mathrm{v}}$ | Yes | 3.60±0.12 | 3.59±0.11 | 3.34±0.11 | 3.48±0.11 |
| S5 | PWG-V/UV-D | $D^{\mathrm{uv}}$ | $D^{\mathrm{v}}$ | Yes | 3.67±0.11 | 3.48±0.11 | 3.29±0.12 | 3.38±0.11 |
| S6 | PWG-V/UV-D | $D^{\mathrm{uv}}$ | $D^{\mathrm{uv}}$ | Yes | 3.77±0.11 | 3.88±0.10 | 3.57±0.11 | 3.34±0.11 |
| **S7** | **PWG-V/UV-D (proposed)** | $D^{\mathrm{v}}$ | $D^{\mathrm{uv}}$ | Yes | **4.11±0.10** | **4.05±0.10** | **4.04±0.10** | **4.08±0.10** |
| R1 | Recordings | - | - | - | 4.63±0.08 | 4.67±0.07 | 4.61±0.08 | 4.64±0.08 |

**Table 3.** MOS test results with 95% confidence intervals: Acoustic features generated from the FastSpeech 2 acoustic model were used to compose the input auxiliary features.

| System | Model | MOS F1 | MOS F2 | MOS M1 | MOS M2 |
|---|---|---|---|---|---|
| S1 | FastSpeech 2 + WaveNet | 3.90±0.11 | 3.81±0.10 | 3.43±0.11 | 3.09±0.10 |
| S2 | FastSpeech 2 + PWG | 3.76±0.11 | 3.62±0.11 | 3.63±0.11 | 3.78±0.10 |
| S3 | FastSpeech 2 + PWG-cGAN-D | 4.02±0.10 | 4.03±0.10 | 4.16±0.10 | 4.06±0.10 |
| **S7** | **FastSpeech 2 + PWG-V/UV-D (proposed)** | **4.20±0.10** | **4.18±0.09** | **4.21±0.09** | **4.31±0.09** |
| R1 | Recordings | 4.63±0.08 | 4.67±0.07 | 4.61±0.08 | 4.64±0.08 |

R. Yamamoto, et al., "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," Proc. ICASSP, 2021, pp. 6039-6043.

# Voicing-aware discriminators

## Evaluation results

**Table 2.** MOS test results with 95% confidence intervals in analysis/synthesis: The speech samples were generated using the acoustic features extracted from the recorded speech. PWG denotes Parallel WaveGAN for short. Note that systems S2 and S3 used $D^{\mathrm{v}}$ as the primary discriminator. All the models were trained in a speaker-independent manner.

| System | Model | Voiced segments | Unvoiced segments | Discriminator conditioning | MOS F1 | MOS F2 | MOS M1 | MOS M2 |
|---|---|---|---|---|---|---|---|---|
| S1 | WaveNet | - | - | - | 3.64±0.12 | 3.83±0.11 | 3.33±0.12 | 3.13±0.11 |
| S2 | PWG | - | - | - | 3.61±0.11 | 3.55±0.11 | 3.57±0.12 | 3.61±0.11 |
| S3 | PWG-cGAN-D | - | - | Yes | 4.04±0.10 | 3.95±0.10 | 3.91±0.11 | 3.97±0.10 |
| S4 | PWG-V/UV-D | $D^{\mathrm{v}}$ | $D^{\mathrm{v}}$ | Yes | 3.60±0.12 | 3.59±0.11 | 3.34±0.11 | 3.48±0.11 |
| S5 | PWG-V/UV-D | $D^{\mathrm{uv}}$ | $D^{\mathrm{v}}$ | Yes | 3.67±0.11 | 3.48±0.11 | 3.29±0.12 | 3.38±0.11 |
| S6 | PWG-V/UV-D | $D^{\mathrm{uv}}$ | $D^{\mathrm{uv}}$ | Yes | 3.77±0.11 | 3.88±0.10 | 3.57±0.11 | 3.34±0.11 |
| **S7** | **PWG-V/UV-D (proposed)** | $D^{\mathrm{v}}$ | $D^{\mathrm{uv}}$ | Yes | **4.11±0.10** | **4.05±0.10** | **4.04±0.10** | **4.08±0.10** |
| R1 | Recordings | - | - | - | 4.63±0.08 | 4.67±0.07 | 4.61±0.08 | 4.64±0.08 |

**Table 3.** MOS test results with 95% confidence intervals: Acoustic features generated from the FastSpeech 2 acoustic model were used to compose the input auxiliary features.

| System | Model | MOS F1 | MOS F2 | MOS M1 | MOS M2 |
|---|---|---|---|---|---|
| S1 | FastSpeech 2 + WaveNet | 3.90±0.11 | 3.81±0.10 | 3.43±0.11 | 3.09±0.10 |
| S2 | FastSpeech 2 + PWG | 3.76±0.11 | 3.62±0.11 | 3.63±0.11 | 3.78±0.10 |
| S3 | FastSpeech 2 + PWG-cGAN-D | 4.02±0.10 | 4.03±0.10 | 4.16±0.10 | 4.06±0.10 |
| **S7** | **FastSpeech 2 + PWG-V/UV-D (proposed)** | **4.20±0.10** | **4.18±0.09** | **4.21±0.09** | **4.31±0.09** |
| R1 | Recordings | 4.63±0.08 | 4.67±0.07 | 4.61±0.08 | 4.64±0.08 |

R. Yamamoto, et al., "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," Proc. ICASSP, 2021, pp. 6039-6043.

# Voicing-aware discriminators

Demo samples

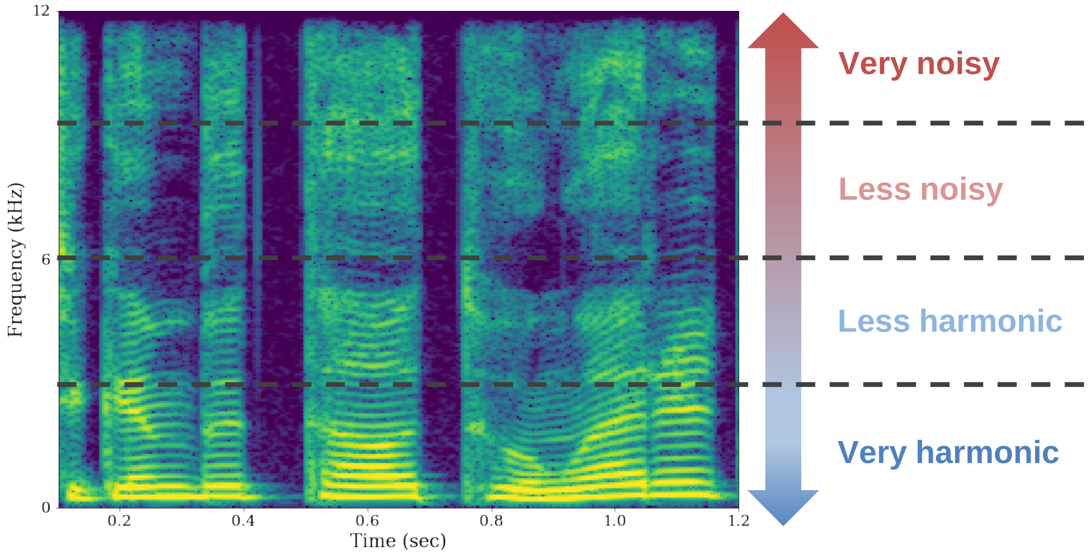
R. Yamamoto, et al., “Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators,” Proc. ICASSP, 2021, pp. 6039-6043.

# Parallel waveform synthesis

Toward high-quality synthesis: Harmonic/noise generators

# Harmonic/noise generators

## Harmonicity analysis in the frequency domain




Low frequency region

Harmonic characteristics > Noise characteristics

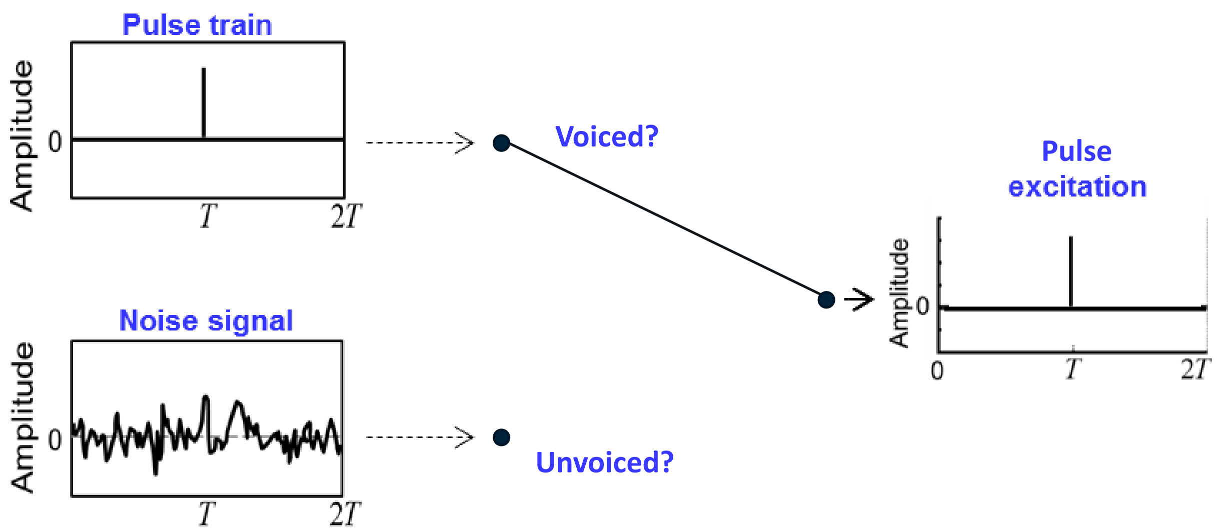
High frequency region

Harmonic characteristics < Noise characteristics

M.-J. Hwang, et al., “High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model,” Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Harmonic/noise generators

## Parametric LPC vocoder (binary decision)




Low frequency region

Harmonic characteristics > Noise characteristics

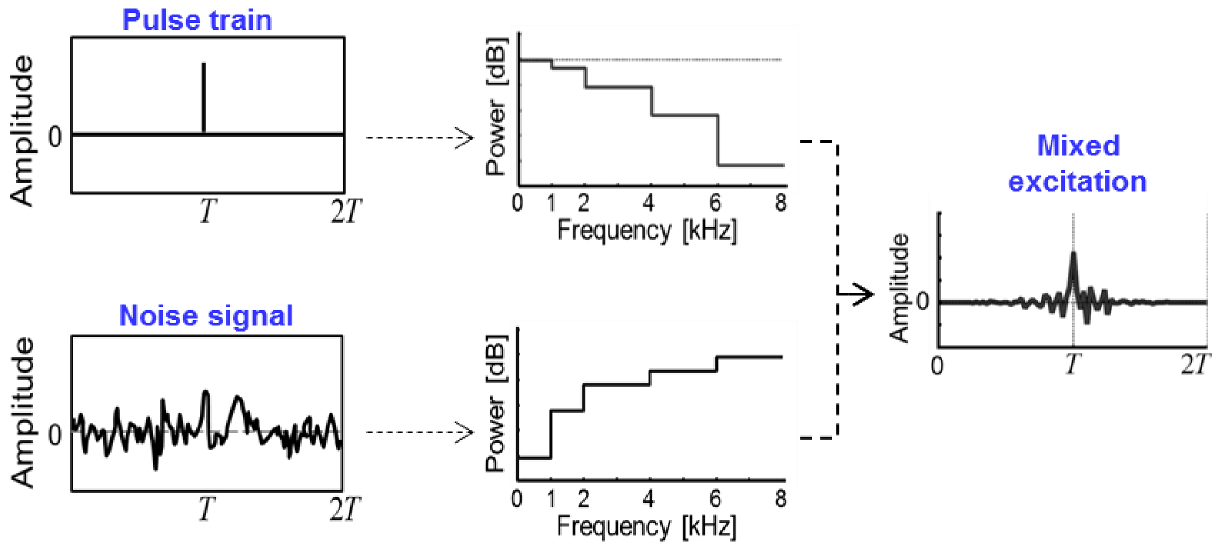
High frequency region

Harmonic characteristics < Noise characteristics

M.-J. Hwang, et al., “High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model,” Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Harmonic/noise generators

## Mixed excitation-based parametric vocoder




Low frequency region

Harmonic characteristics > Noise characteristics

High frequency region

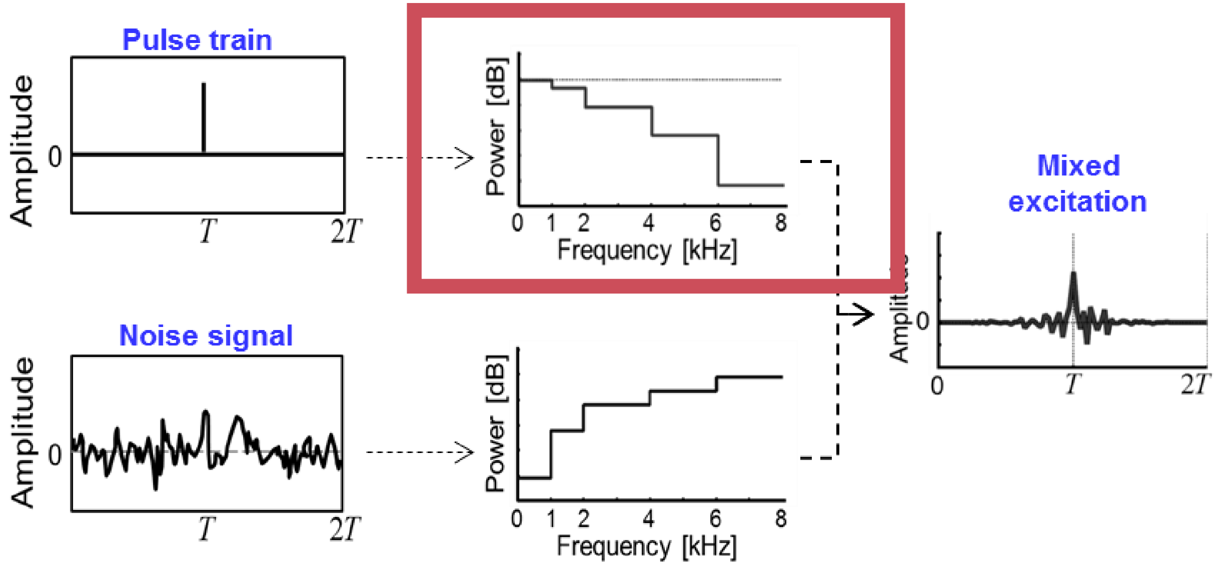
Harmonic characteristics < Noise characteristics

M.-J. Hwang, et al., “High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model,” Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Harmonic/noise generators

## Mixed excitation-based parametric vocoder

How periodic? → Harmonicity (ex. MELP and MBE vocoders)




Low frequency region

Harmonic characteristics > Noise characteristics

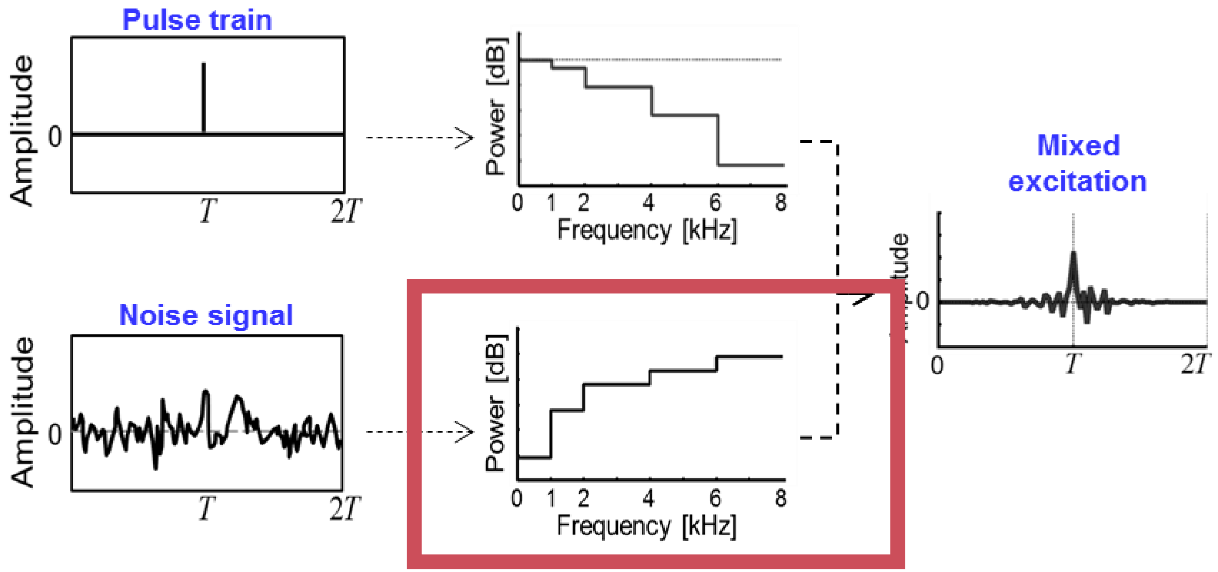
High frequency region

Harmonic characteristics < Noise characteristics

M.-J. Hwang, et al., "High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model," Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Harmonic/noise generators

## Mixed excitation-based parametric vocoder




How aperiodic? → aperiodicity (ex. STRAIGHT and WORLD vocoders)

Low frequency region

Harmonic characteristics > Noise characteristics

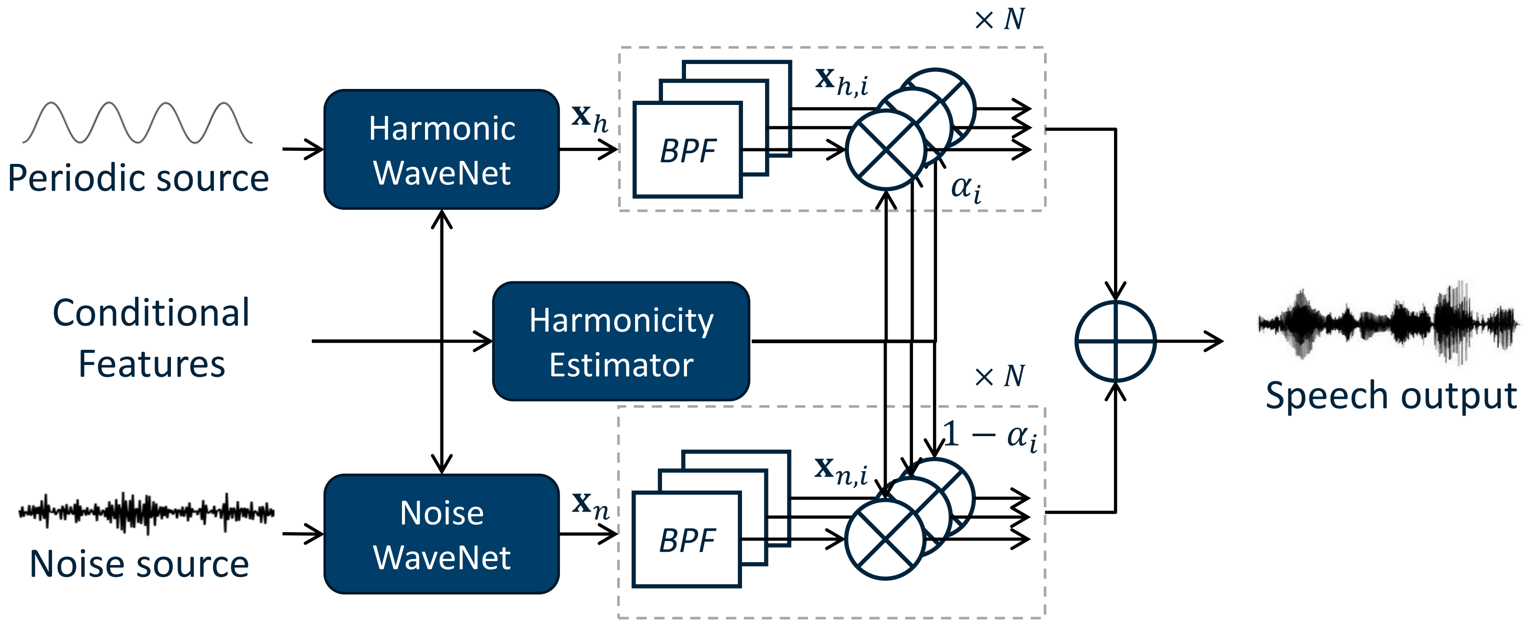
High frequency region

Harmonic characteristics < Noise characteristics

M.-J. Hwang, et al., “High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model,” Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Harmonic/noise generators

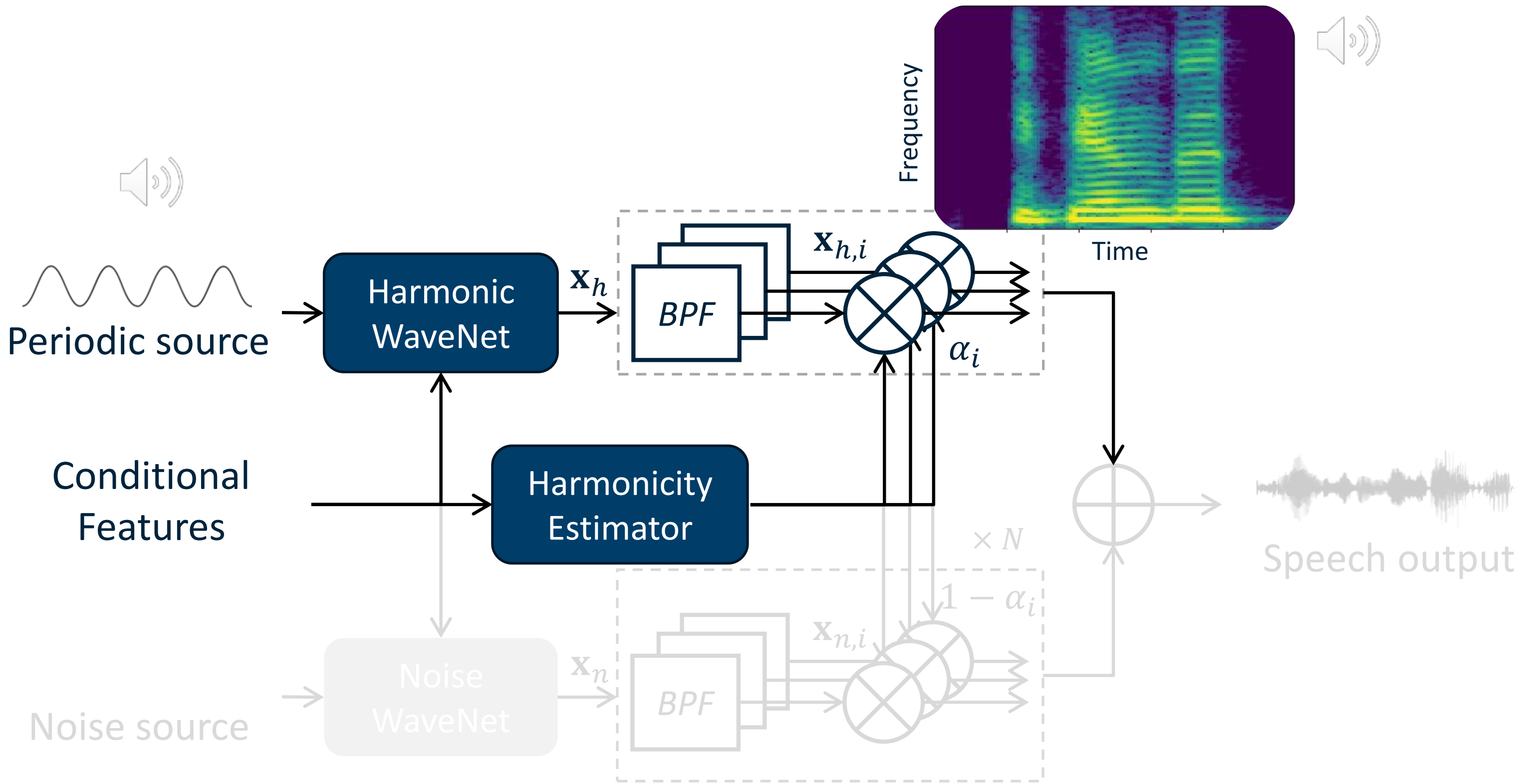
## Model architecture




M.-J. Hwang, et al., "High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model," Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Harmonic/noise generators

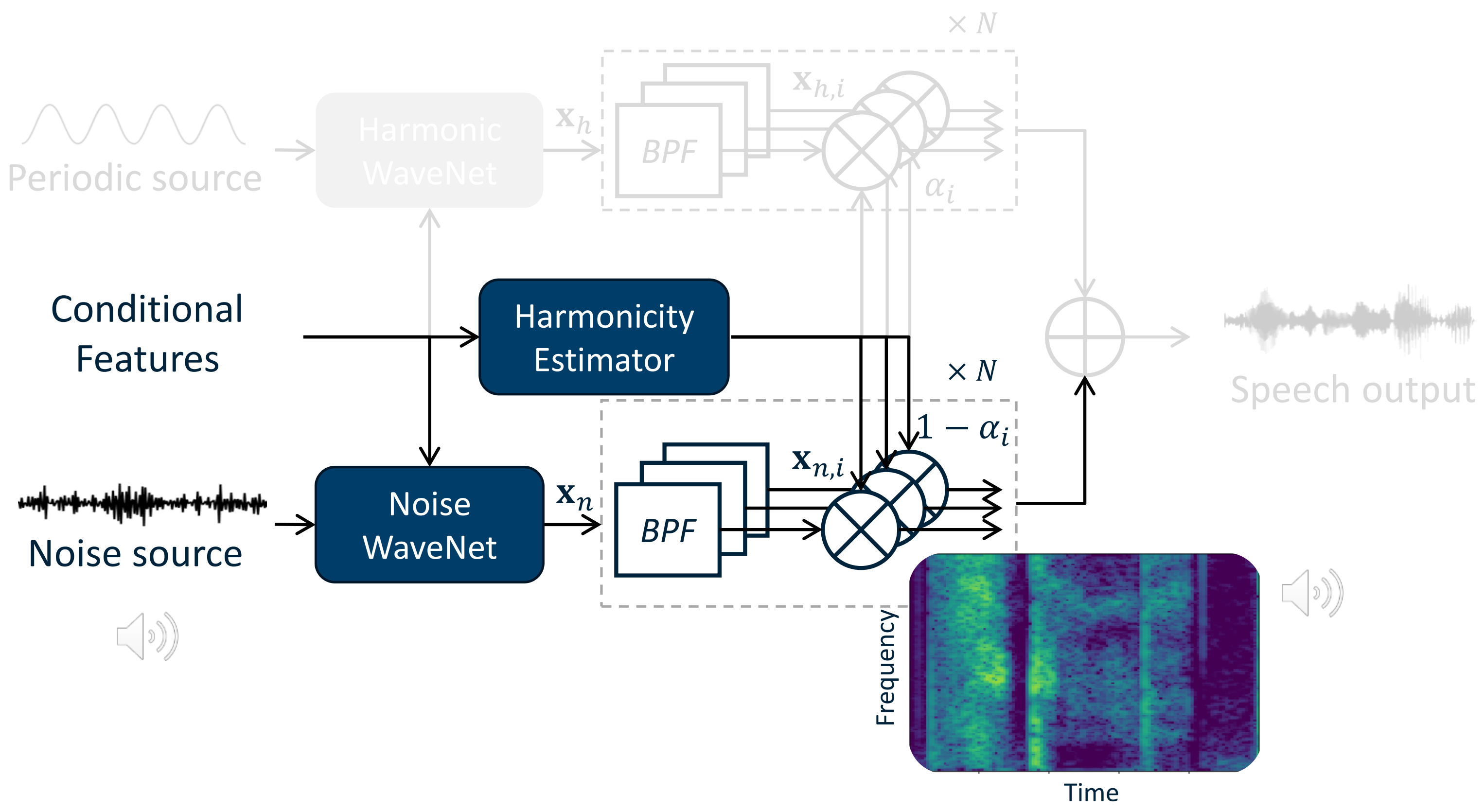
## Model architecture




M.-J. Hwang, et al., "High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model," Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Harmonic/noise generators

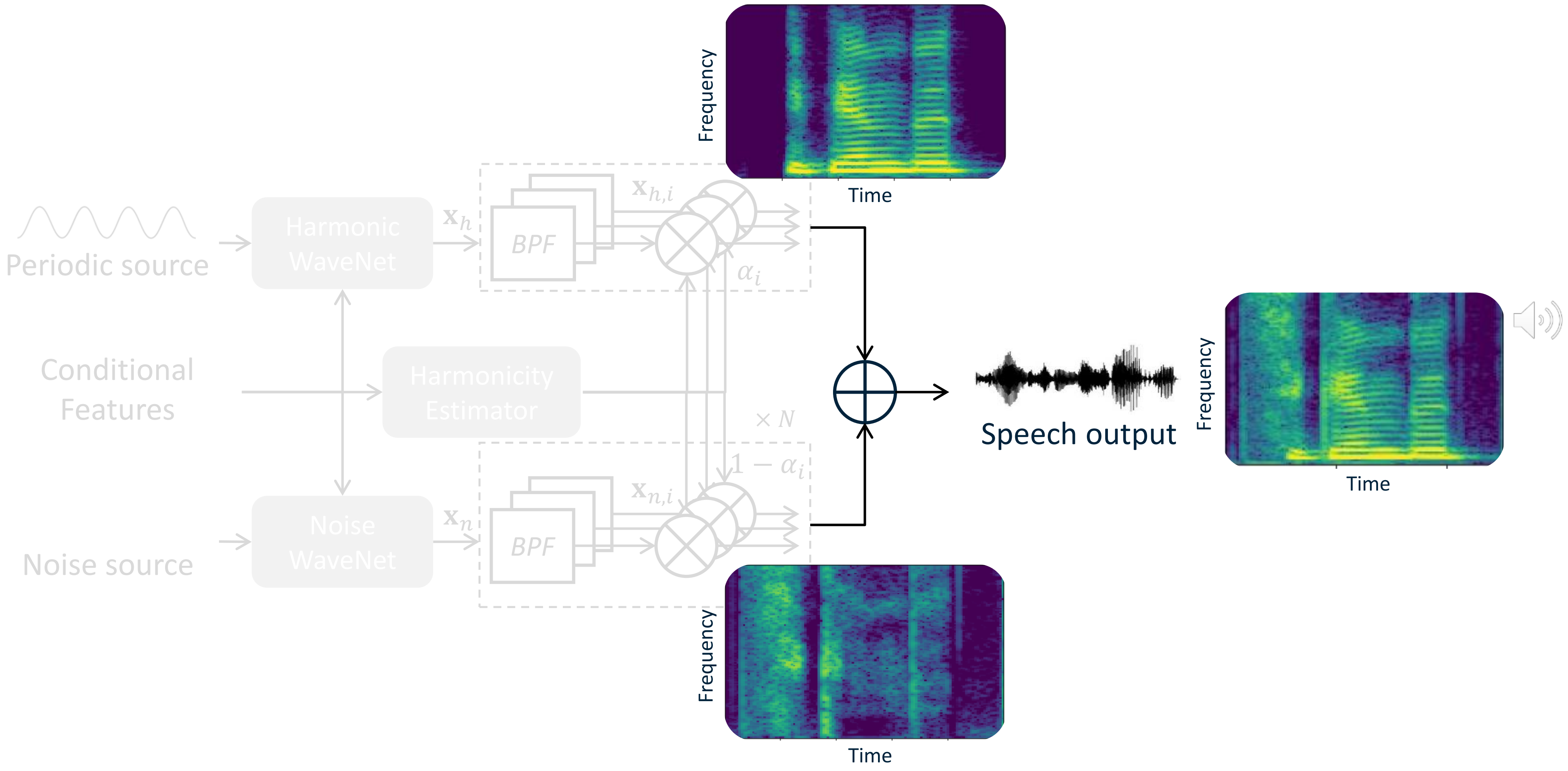
## Model architecture




M.-J. Hwang, et al., "High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model," Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Harmonic/noise generators

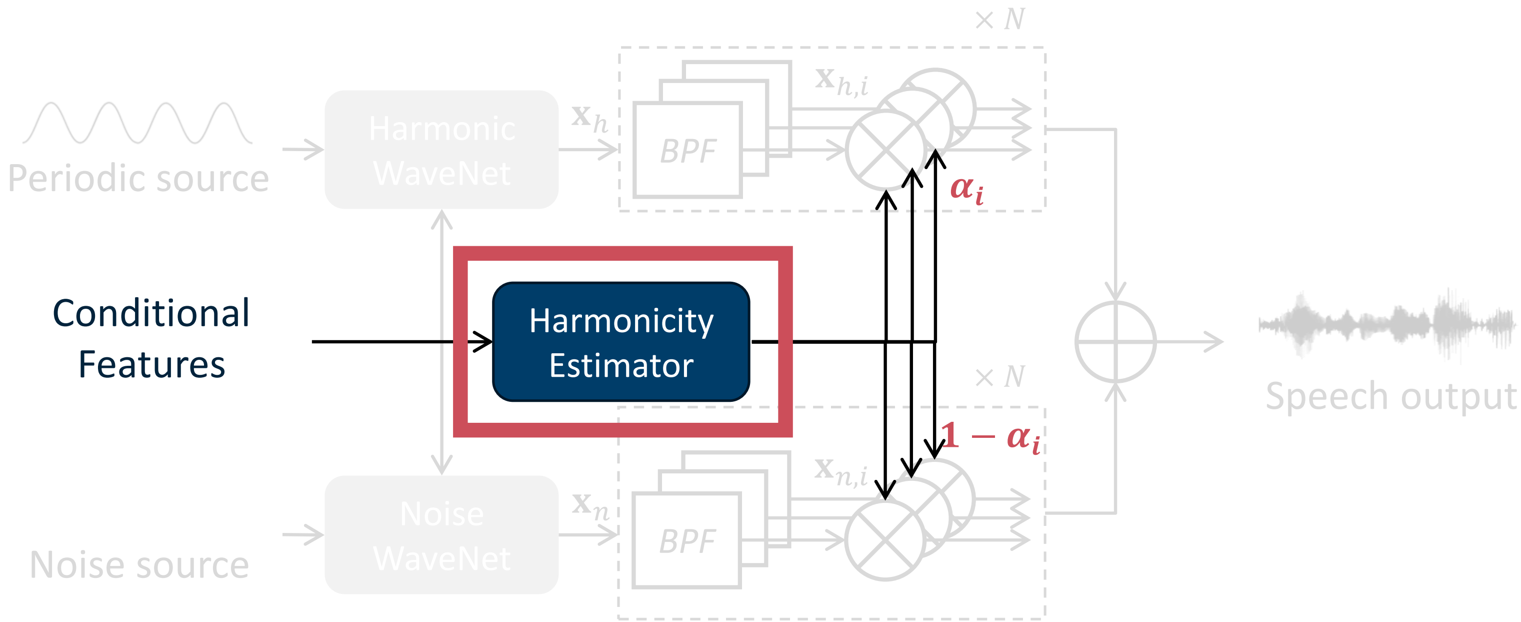
### Model architecture




M.-J. Hwang, et al., "High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model," Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Harmonic/noise generators

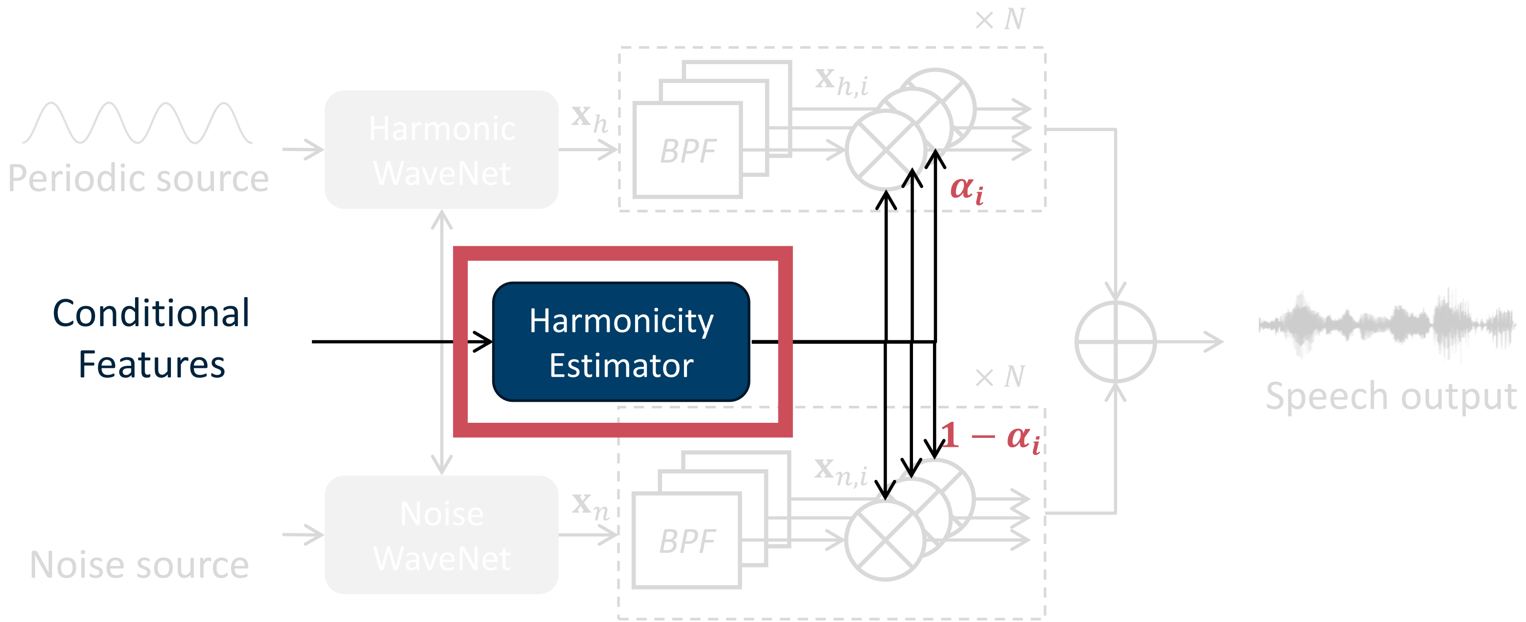
## Model architecture




M.-J. Hwang, et al., "High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model," Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Harmonic/noise generators

**Model architecture**

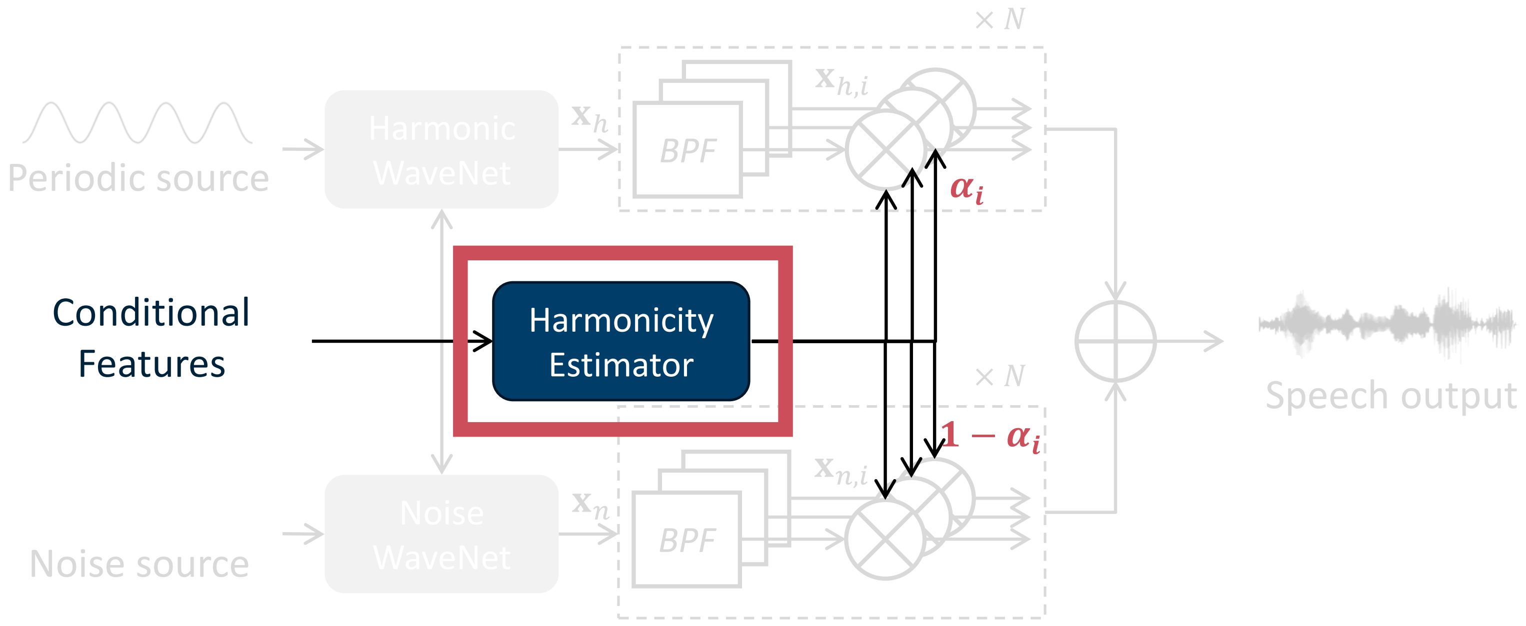



Parametric vocoders: Harmonicity has been estimated by rule-based analysis methods

M.-J. Hwang, et al., "High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model," Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Harmonic/noise generators

## Model architecture




Parametric vocoders: Harmonicity has been estimated by rule-based analysis methods
Alternatively, we design learnable harmonicities optimized CNN blocks with input condition

M.-J. Hwang, et al., “High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model,” Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Harmonic/noise generators

## Evaluation results

Table 1. *The model size, inference speed, and MOS results with 95% confidence intervals: Acoustic features extracted from the recorded speech signal were used to compose the input acoustic features. The MOS results for highest score is in bold font.*

| Label | Model | Use of HN model | Input signals for H-WaveNet | Type of HN model | Model size (M) | Inference speed | MOS |
|---|---|---|---|---|---|---|---|
| S1 | WaveNet [21] | – | – | – | 3.81 | $0.34 \times 10^{-2}$ | $4.22 \pm 0.12$ |
| S2 | PWG [7] | – | – | – | 0.94 | 50.38 | $3.46 \pm 0.37$ |
| S3 | HN-PWG w/o noise [16] | Yes | Sine + V/UV | Full-band | 0.94 | 47.91 | $4.02 \pm 0.14$ |
| S4 | HN-PWG | Yes | Sine + noise + V/UV | Full-band | 0.94 | 47.93 | $4.18 \pm 0.15$ |
| **S5** | **Multi-band HN-PWG** | **Yes** | **Sine + noise + V/UV** | **Multi-band** | **0.99** | **47.87** | $\mathbf{4.29 \pm 0.12}$ |
| S6 | Recordings | – | – | – | – | – | $4.41 \pm 0.12$ |

S$i$: $i^{\text{th}}$ system; HN: harmonic-plus-noise; PWG: Parallel WaveGAN; H-WaveNet: harmonic WaveNet; V/UV: voicing flags upsampled from frame-level to sample-level. Note that inference speed, $k$, indicates that a system was able to generate waveforms $k$ times faster than real-time. This evaluation was conducted on a server with a single NVIDIA Tesla V100 GPU.

Table 2. *Subjective MOS test results with 95% confidence intervals for the TTS systems with respect to the different vocoding models. The MOS results for highest score is in bold font.*

| Label | Model | MOS |
|---|---|---|
| S-T1 | WaveNet [21] | $4.03 \pm 0.19$ |
| S-T2 | PWG [7] | $3.56 \pm 0.28$ |
| S-T3 | HN-PWG w/o noise | $2.60 \pm 0.22$ |
| S-T4 | HN-PWG | $4.01 \pm 0.17$ |
| **S-T5** | **Multi-band HN-PWG** | $\mathbf{4.03 \pm 0.16}$ |
| S6 | Recordings | $4.41 \pm 0.12$ |

S-T$i$: $i^{\text{th}}$ system that generates speech waveform from the acoustic features predicted by TTS model.

Acoustic model: Tacotron 2

M.-J. Hwang, et al., “High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model,” Proc. INTERSPEECH, 2021, pp. 2227-2231.

# Parallel waveform synthesis

## Summary

# Summary

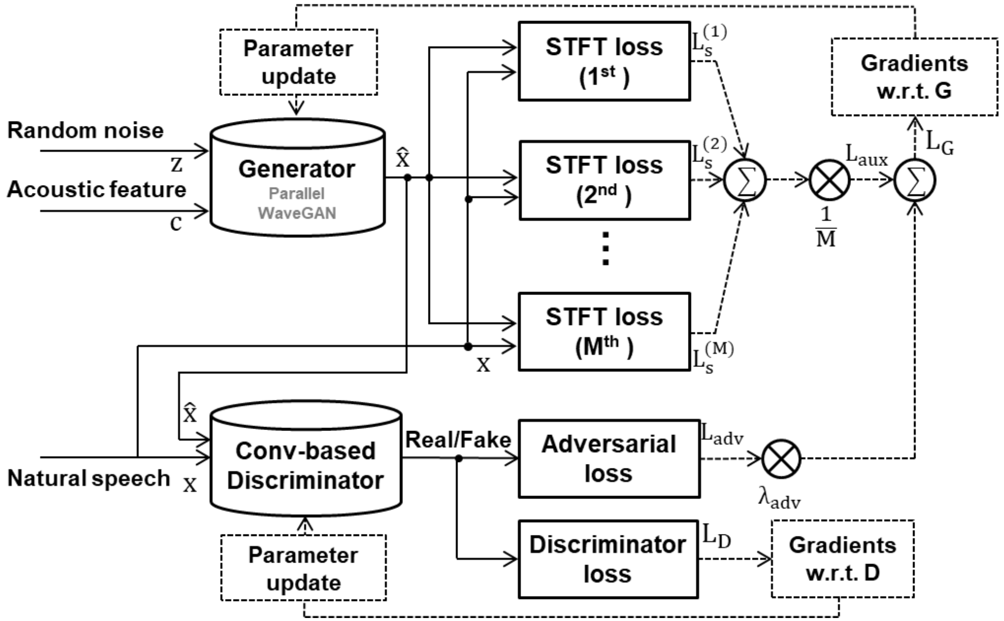
## PARALLEL WAVEGAN: A FAST WAVEFORM GENERATION MODEL BASED ON GENERATIVE ADVERSARIAL NETWORKS WITH MULTI-RESOLUTION SPECTROGRAM

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

[1]LINE Corp., Tokyo, Japan.
[2]NAVER Corp., Seongnam, Korea

### ABSTRACT

We propose Parallel WaveGAN, a distillation-free, fast, and small-footprint waveform generation method using a generative adversarial network. In the proposed method, a non-autoregressive WaveNet is trained by jointly optimizing multi-resolution spectrogram and adversarial loss functions, which can effectively capture the time-frequency distribution of the realistic speech waveform. As our method does not require density distillation used in the conventional teacher-student framework, the entire model can be easily trained. Furthermore, our model is able to generate high-fidelity speech even with its compact architecture. In particular, the proposed Parallel WaveGAN has only 1.44 M parameters and can generate 24 kHz speech waveform 28.68 times faster than real-time on a single GPU environment. Perceptual listening test results verify that our proposed method achieves 4.16 mean opinion score within a Transformer-based text-to-speech framework, which is comparative to the best distillation-based Parallel WaveNet system.



# Summary

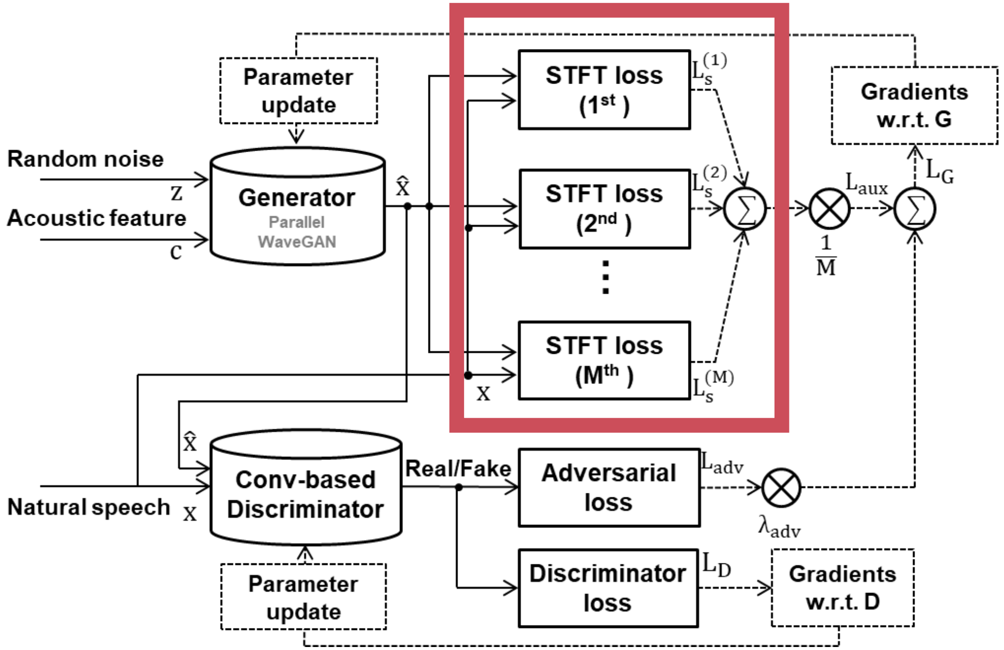
## IMPROVED PARALLEL WAVEGAN VOCODER WITH PERCEPTUALLY WEIGHTED SPECTROGRAM LOSS

*Eunwoo Song*[1], *Ryuichi Yamamoto*[2], *Min-Jae Hwang*[3], *Jin-Seob Kim*[1], *Ohsung Kwon*[1], *Jae-Min Kim*[1]

[1]NAVER Corp., Seongnam, Korea
[2]LINE Corp., Tokyo, Japan
[3]Search Solutions Inc., Seongnam, Korea

### ABSTRACT

This paper proposes a spectral-domain perceptual weighting technique for Parallel WaveGAN-based text-to-speech (TTS) systems. The recently proposed Parallel WaveGAN vocoder successfully generates waveform sequences using a fast non-autoregressive WaveNet model. By employing multi-resolution short-time Fourier transform (MR-STFT) criteria with a generative adversarial network, the light-weight convolutional networks can be effectively trained without any distillation process. To further improve the vocoding performance, we propose the application of frequency-dependent weighting to the MR-STFT loss function. The proposed method penalizes perceptually-sensitive errors in the frequency domain; thus, the model is optimized toward reducing auditory noise in the synthesized speech. Subjective listening test results demonstrate that our proposed method achieves 4.21 and 4.26 TTS mean opinion scores for female and male Korean speakers, respectively.



# Summary

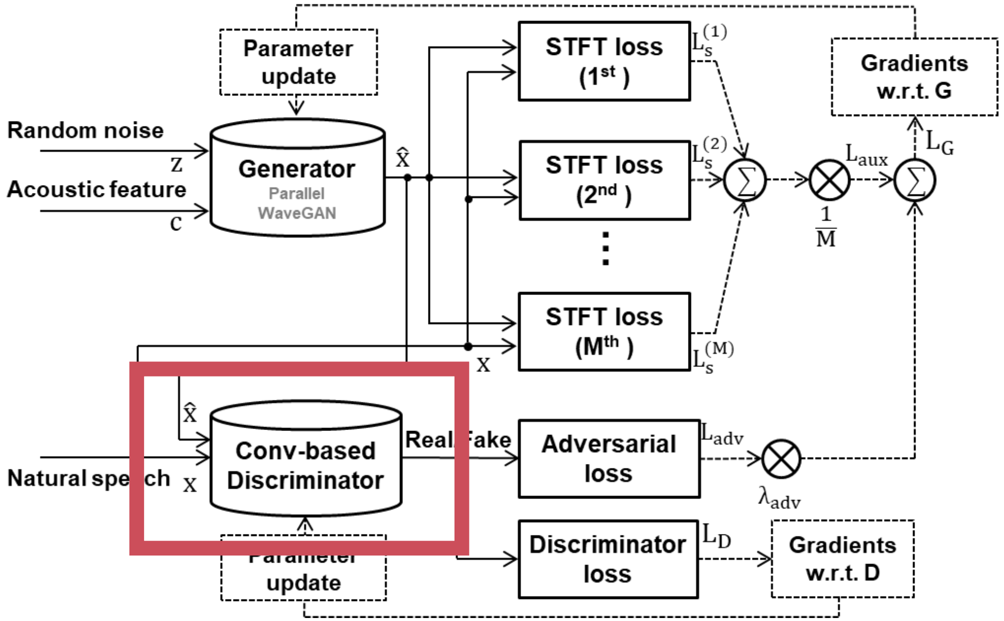
**PARALLEL WAVEFORM SYNTHESIS BASED ON GENERATIVE ADVERSARIAL NETWORKS WITH VOICING-AWARE CONDITIONAL DISCRIMINATORS**

*Ryuichi Yamamoto*[1], *Eunwoo Song*[2], *Min-Jae Hwang*[3] *and Jae-Min Kim*[2]

[1]LINE Corp., Tokyo, Japan
[2]NAVER Corp., Seongnam, Korea
[3]Search Solutions Inc., Seongnam, Korea

**ABSTRACT**

This paper proposes voicing-aware conditional discriminators for Parallel WaveGAN-based waveform synthesis systems. In this framework, we adopt a projection-based conditioning method that can significantly improve the discriminator's performance. Furthermore, the conventional discriminator is separated into two waveform discriminators for modeling voiced and unvoiced speech. As each discriminator learns the distinctive characteristics of the harmonic and noise components, respectively, the adversarial training process becomes more efficient, allowing the generator to produce more realistic speech waveforms. Subjective test results demonstrate the superiority of the proposed method over the conventional Parallel WaveGAN and WaveNet systems. In particular, our speaker-independently trained model within a FastSpeech 2 based text-to-speech framework achieves the mean opinion scores of 4.20, 4.18, 4.21, and 4.31 for four Japanese speakers, respectively.



# Summary

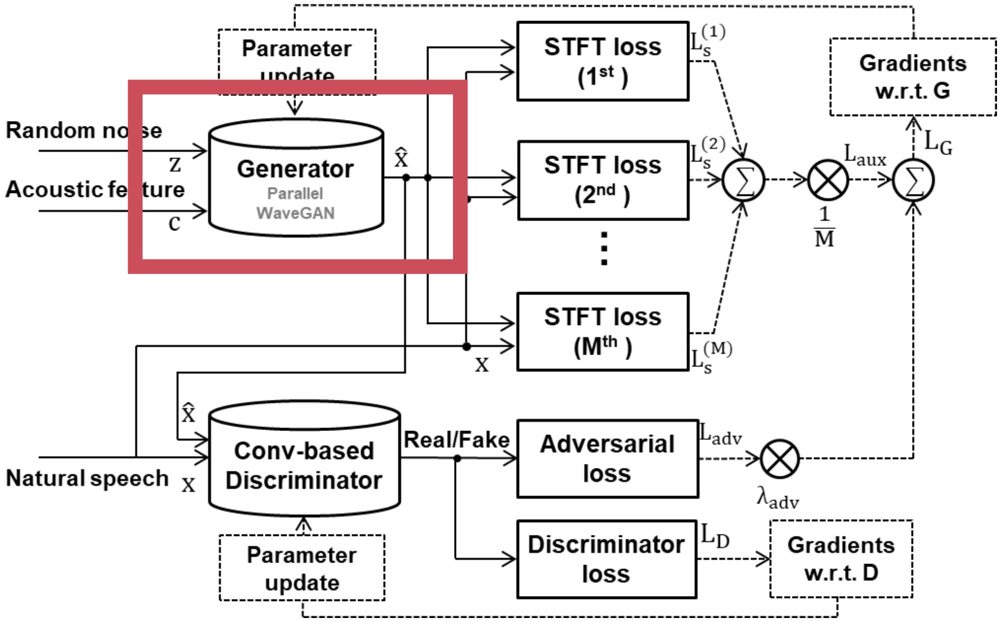
## High-fidelity Parallel WaveGAN with Multi-band Harmonic-plus-Noise Model

*Min-Jae Hwang[1]*, Ryuichi Yamamoto[2]*, Eunwoo Song[3] and Jae-Min Kim[3]*

[1]Search Solutions Inc., Seongnam, Korea
[2]LINE Corp.,Tokyo, Japan
[3]NAVER Corp., Seongnam, Korea

### Abstract

This paper proposes a multi-band harmonic-plus-noise (HN) Parallel WaveGAN (PWG) vocoder. To generate a high-fidelity speech signal, it is important to well-reflect the harmonic-noise characteristics of the speech waveform in the time-frequency domain. However, it is difficult for the conventional PWG model to accurately match this condition, as its single generator inefficiently represents the complicated nature of harmonic-noise structures. In the proposed method, the HN WaveNet models are employed to overcome this limitation, which enable the separate generation of the harmonic and noise components of speech signals from the pitch-dependent sine wave and Gaussian noise sources, respectively. Then, the energy ratios between harmonic and noise components in multiple frequency bands (i.e., subband harmonicities) are predicted by an additional harmonicity estimator. Weighted by the estimated harmonicities, the gain of harmonic and noise components in each subband is adjusted, and finally mixed together to compose the full-band speech signal. Subjective evaluation results showed that the proposed method significantly improved the perceptual quality of the synthesized speech.



# Q / A

gregorio.song@gmail.com