Clova

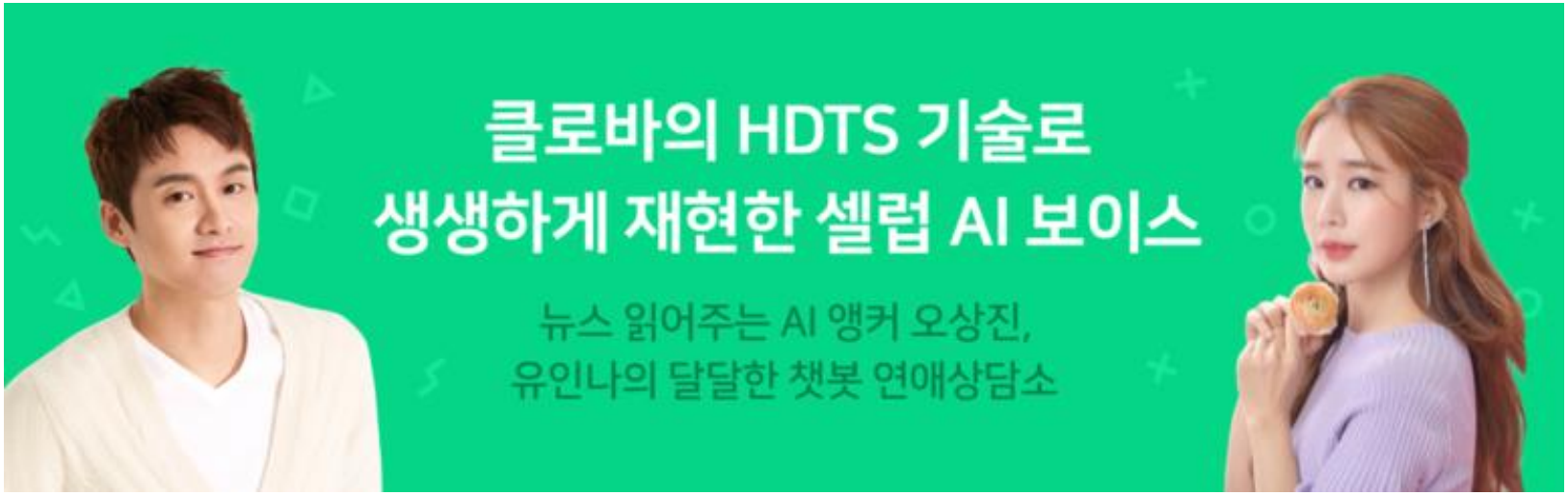
Parallel WaveGAN

빠르고 안정적인 WaveNet 음성 합성 모델 만들기

송은우 / HDTS

Introduction

Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.



Introduction

Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.

DNN TTS = Acoustic model + Vocoder

Introduction

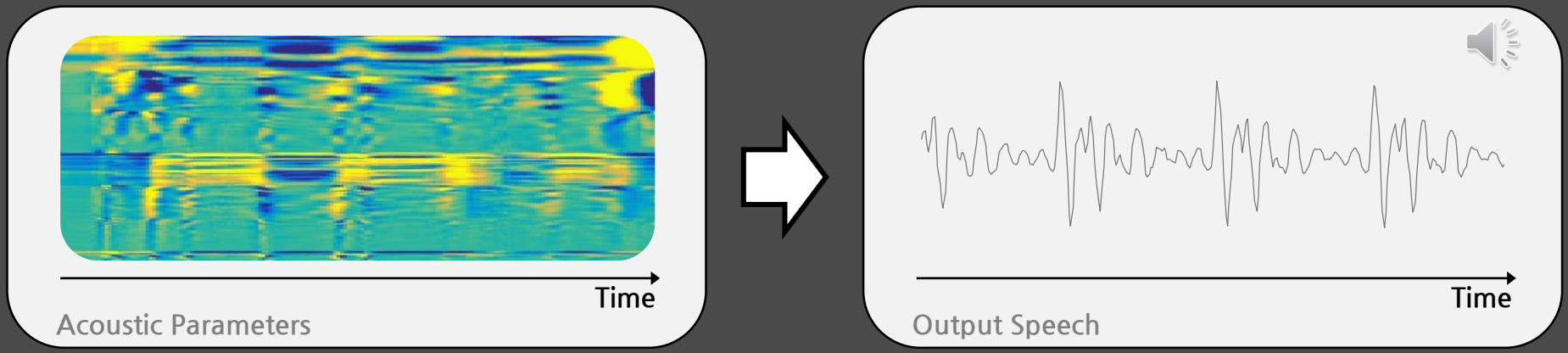
Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.

톤의 높낮이, 음색, 어조, 강세 등
텍스트에서 **Acoustic Parameter** 를 추정

Introduction

Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.

Acoustic Parameter 에서 음성 신호를 추정



Introduction

Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.

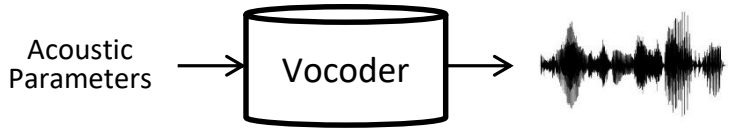
Acoustic Parameter 에서 음성 신호를 추정

본 발표에서는 TTS 엔진의 핵심 요소인 **Vocoder** 기술을 정리하고,
빠르고 안정적인 **Parallel WaveGAN** 모델을 소개하고자 합니다.

Vocoder: Overview

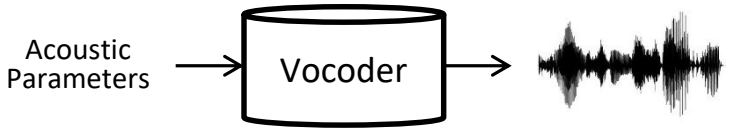
Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.



Vocoder: Overview

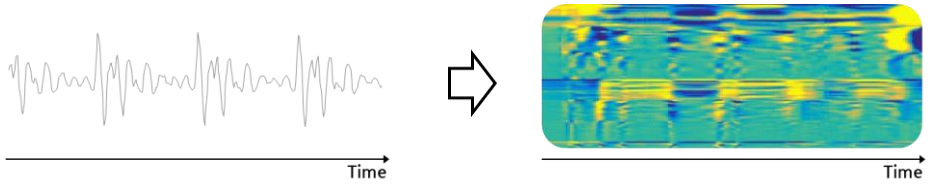
Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.



Acoustic Parameter ..?

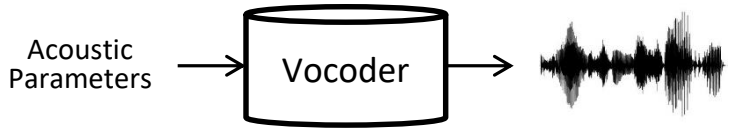
톤의 높낮이, 음색, 어조, 강세 등 음성의 특징을 나타내는 파라미터

(ex. F0, Spectrum, v/uv, ...)



Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.

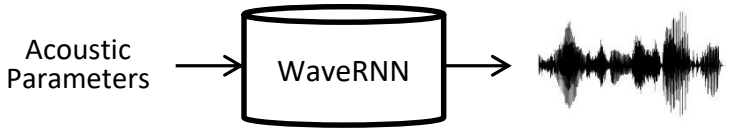


어떤 모델을 Vocoder 에 사용하냐에 따라 종류가 정해집니다.

Clova

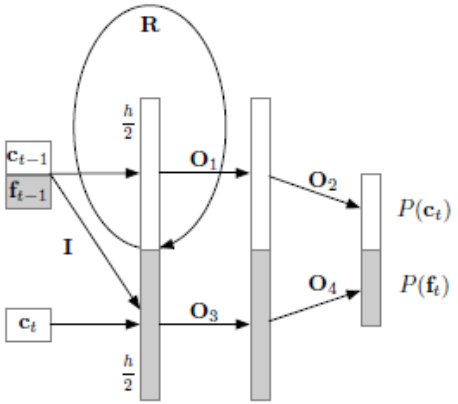
Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.



어떤 모델을 Vocoder 에 사용하냐에 따라 종류가 정해집니다.

RNN 기반의 **WaveRNN**



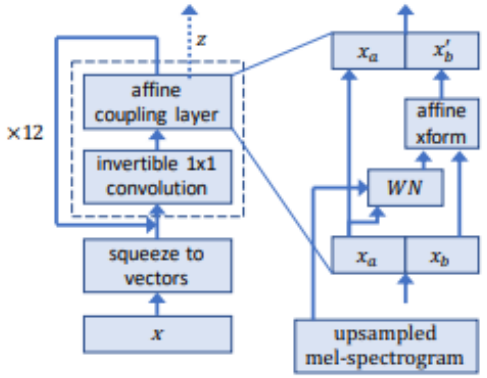
출처: N. Kalchbrenner, et al., “Efficient neural audio synthesis,” *arXiv:1802.08435*, 2018.

Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.

어떤 모델을 Vocoder 에 사용하냐에 따라 종류가 정해집니다.

Flow 기반의 **WaveGlow**

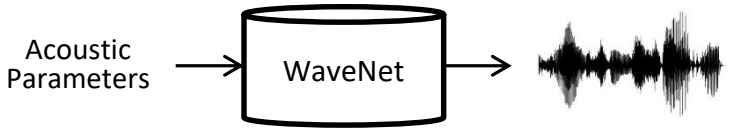


출처: R. Prenger, et al., "WaveGlow: A flow-based generative network for speech synthesis." in *Proc. ICASSP*, 2019.

Clova

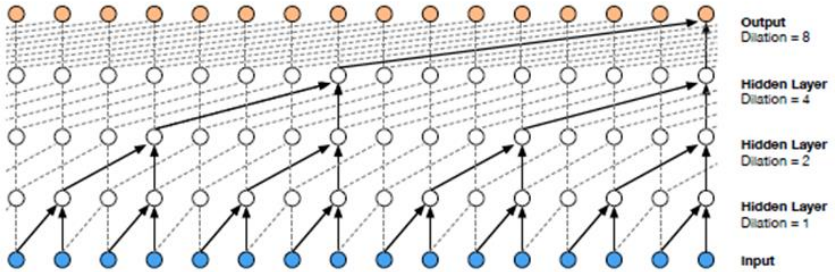
Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.



어떤 모델을 Vocoder 에 사용하냐에 따라 종류가 정해집니다.

CNN 기반의 **WaveNet**



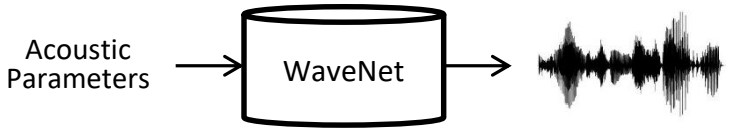
$$p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^{T} p(x_t|x_1, \cdots, x_{t-1}, \mathbf{h})$$

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

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

Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.



어떤 모델을 Vocoder 에 사용하냐에 따라 종류가 정해집니다.

CNN 기반의 **WaveNet**

현재 음성 신호를 예측할 때 과거 음성 신호를 함께 사용합니다.
이러한 방법을 **Autoregressive Model** 라고 정의합니다.

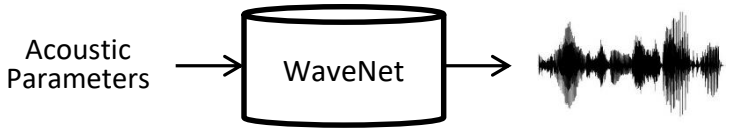
$$p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^{T} p(x_t|x_1, \cdots, x_{t-1}, \mathbf{h})$$

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

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

Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.



어떤 모델을 Vocoder 에 사용하냐에 따라 종류가 정해집니다.

CNN 기반의 **WaveNet**

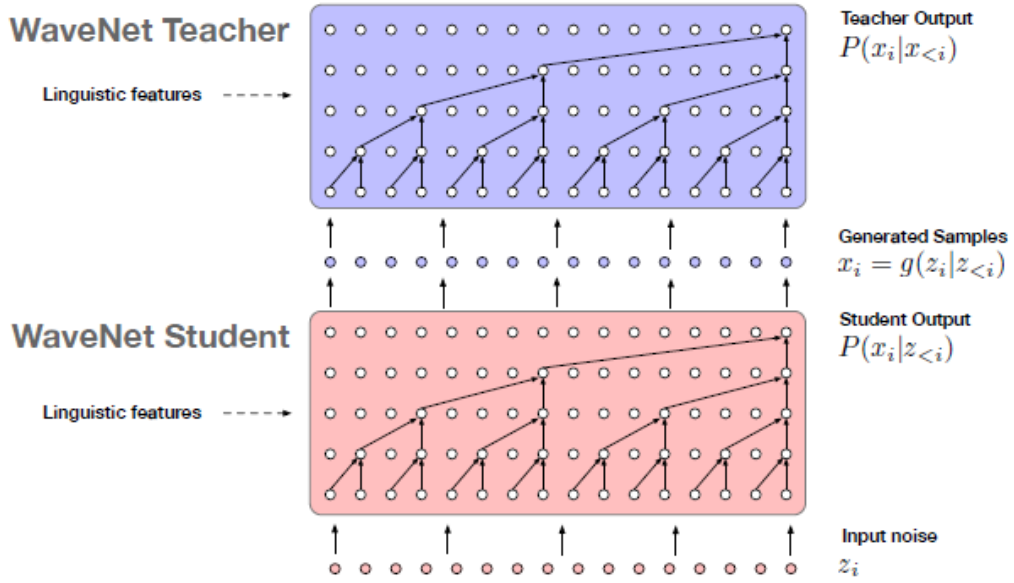
현재 음성 신호를 예측할 때 과거 음성 신호를 함께 사용합니다.
이러한 방법을 **Autoregressive Model** 라고 정의합니다.

Autoregressive Model 은 고품질의 음성을 생성할 수 있으나,
1초 음성을 만들 때 약 5분 정도의 시간이 소요된다는 치명적인 문제가 있습니다.

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

Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.

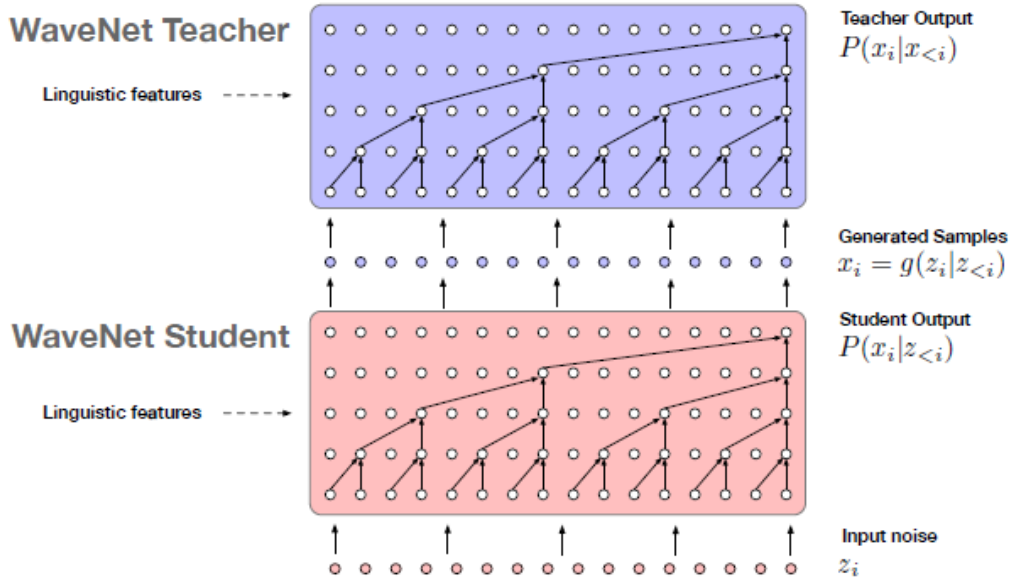


WaveNet 의 속도 문제를 해결하기 위해 제안된 방법이
Non-autoregressive 구조의 **Parallel WaveNet** 입니다.

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

Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.

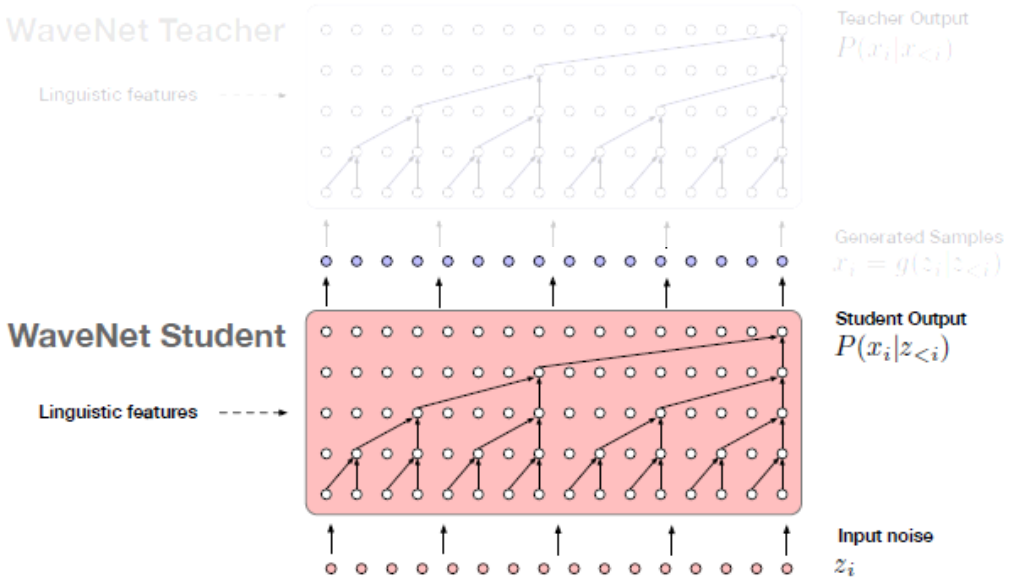


Autoregressive WaveNet (=Teacher) 모델의 확률 분포를
Non-autoregressive Parallel WaveNet (=Student) 모델이 배우도록 훈련합니다.

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

Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.

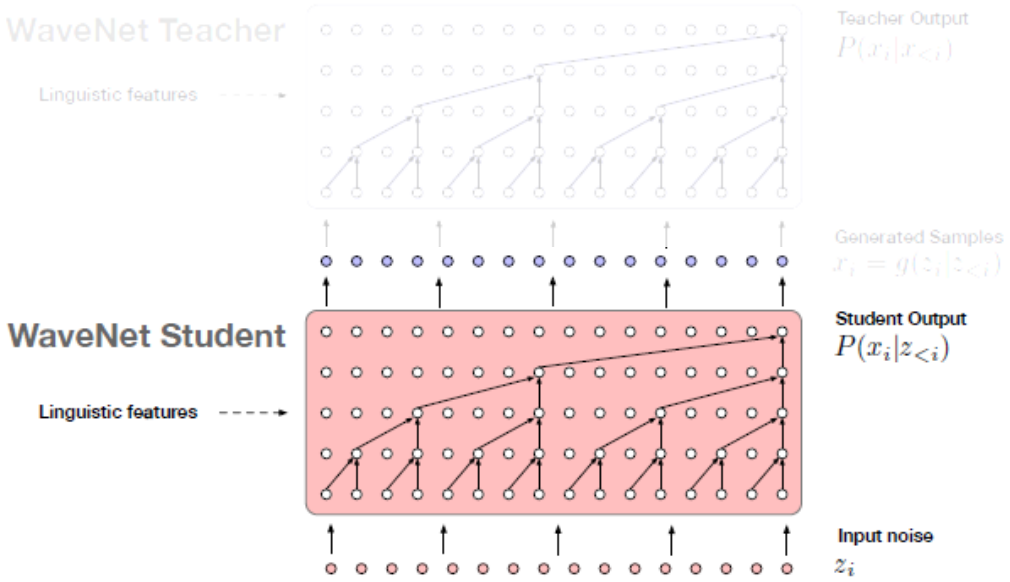


Non-autoregressive Parallel WaveNet 모델은
과거 음성을 사용하지 않으므로, 생성 속도에 제한이 없습니다.
(1초 음성을 약 0.02초 만에 생성 가능)

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

Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.

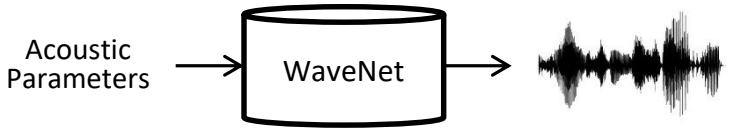


하지만 그만큼 합성음의 품질이 저하된다는 문제가 남아있습니다.

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

Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.



어떤 모델을 Vocoder 에 사용하냐에 따라 종류가 정해집니다.

CNN 기반의 **WaveNet**

Autoregressive

합성음 품질이 좋지만
생성 속도가 느리다

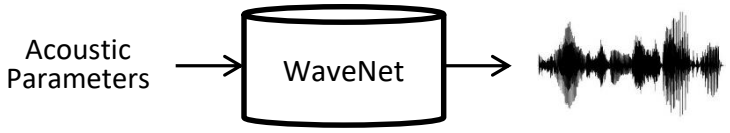
vs

Non-autoregressive

생성 속도가 빠르지만
학습이 어렵고
합성음 품질이 나쁘다

Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.



어떤 모델을 Vocoder 에 사용하냐에 따라 종류가 정해집니다.

CNN 기반의 **WaveNet**

Autoregressive

합성음 품질이 좋지만
생성 속도가 느리다

vs

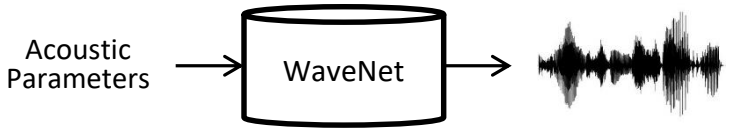
Non-autoregressive

생성 속도가 빠르지만
학습이 어렵고
합성음 품질이 나쁘다

합성음 품질도 좋고, 생성 속도도 빠른 Vocoder 는 없을까

Vocoder: Overview

Vocoder 란 Acoustic Parameter 에서 **음성 신호를 만들어내는** 기술입니다.



어떤 모델을 Vocoder 에 사용하냐에 따라 종류가 정해집니다.

CNN 기반의 **WaveNet**

Autoregressive

합성음 품질이 좋지만
생성 속도가 느리다

vs

Non-autoregressive

생성 속도가 빠르지만
학습이 어렵고
합성음 품질이 나쁘다

Non-autoregressive 의 품질을 높여보자

Vocoder: Parallel WaveNet

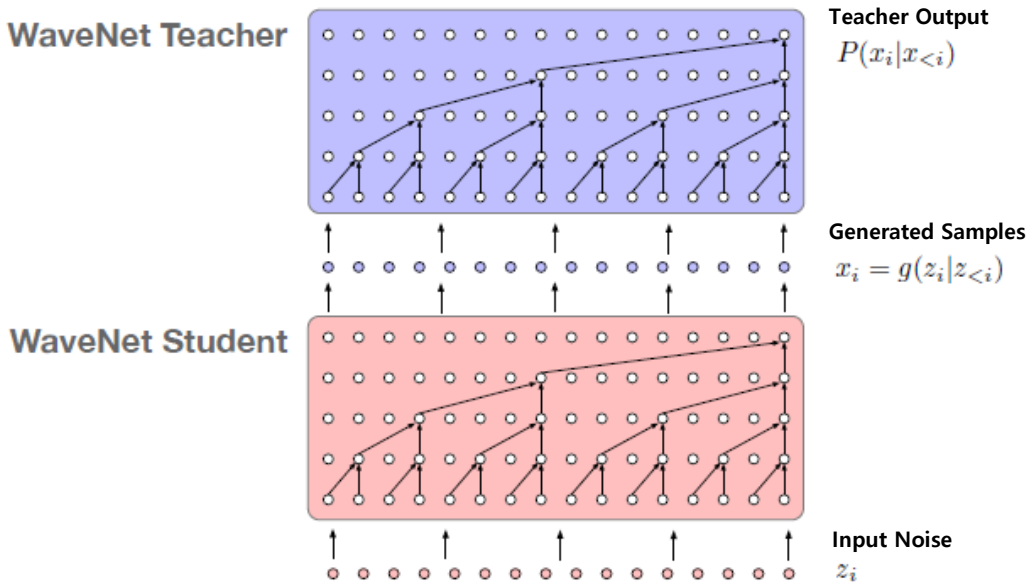
Vocoder: Parallel WaveNet

Vocoder: Parallel WaveGAN

Clova

Vocoder: Parallel WaveGAN

1. Teacher-student 기반의 Probability Distillation 과정을 없애고,



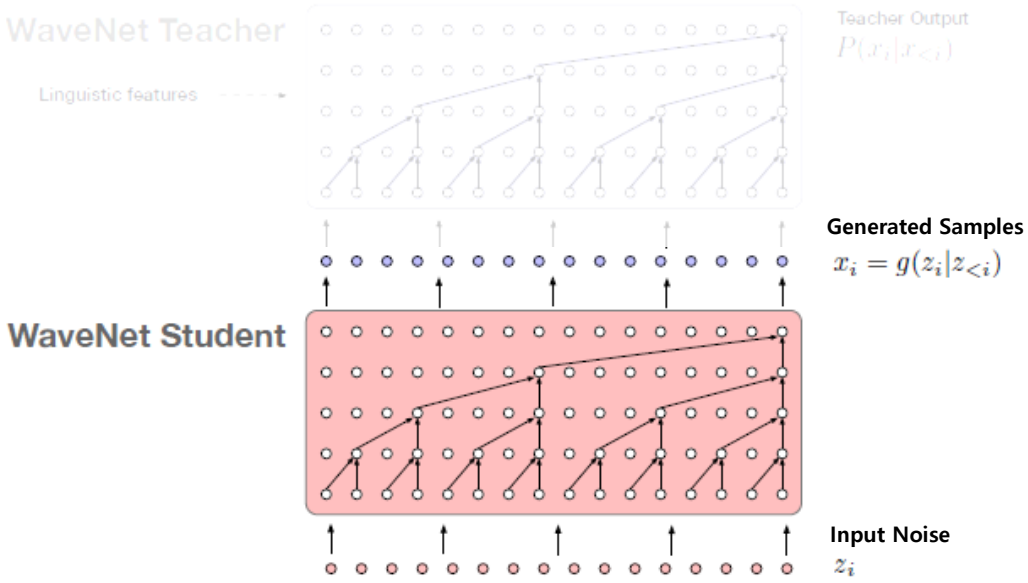
출처: R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

Clova

Vocoder: Parallel WaveGAN

1. Teacher-student 기반의 Probability Distillation 과정을 없애고,

학습이 너무 어려우니까

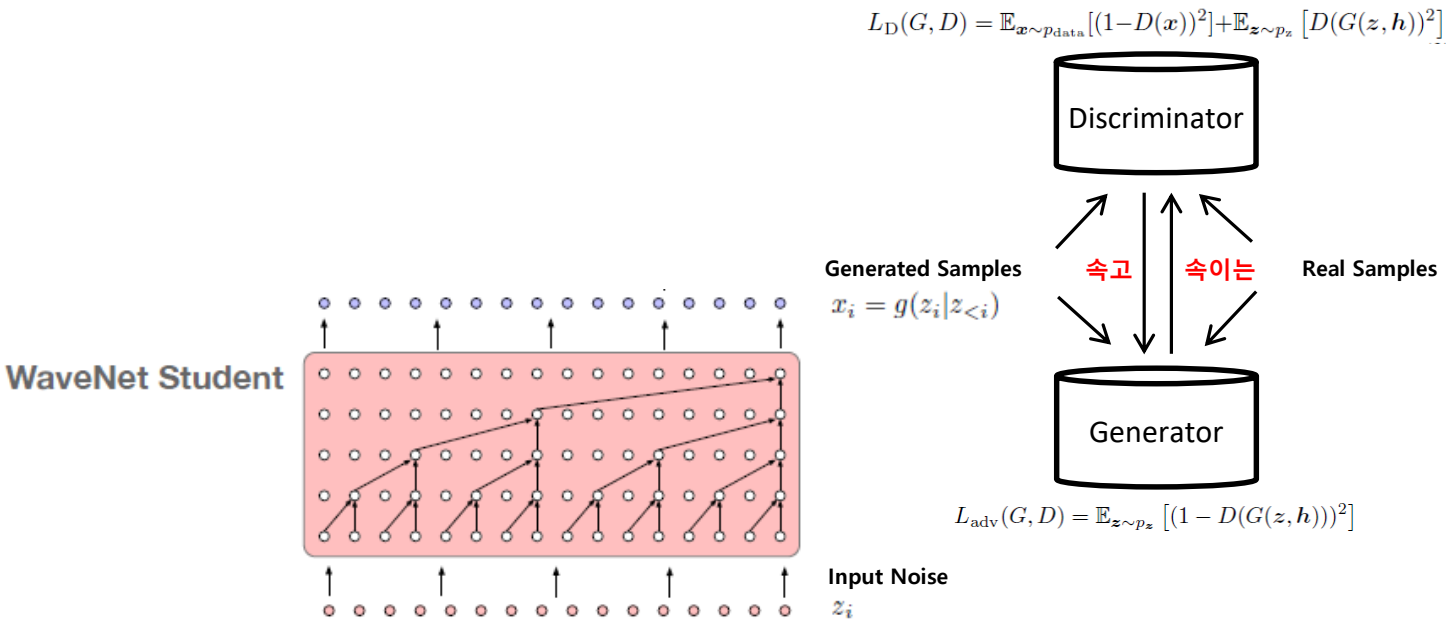


출처: R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

Clova

Vocoder: Parallel WaveGAN

1. Teacher-student 기반의 Probability Distillation 과정을 없애고,
2. Adversarial Training 으로 합성음 품질을 높이고,

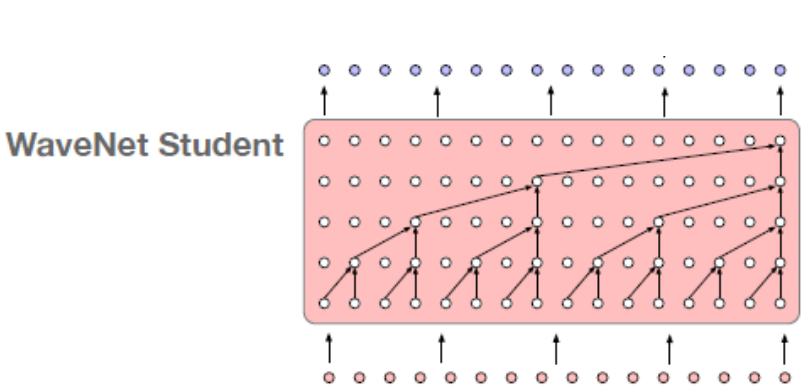


출처: X.Mao, et al.,“Least squares generative adversarial networks,” in *Proc. ICCV*, 2017, pp. 2794-2802.
출처: R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

Clova

Vocoder: Parallel WaveGAN

1. Teacher-student 기반의 Probability Distillation 과정을 없애고,
2. Adversarial Training 으로 합성음 품질을 높이고,
3. Multi-resolution STFT Loss 로 합성음 품질을 더 높이고,



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

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

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

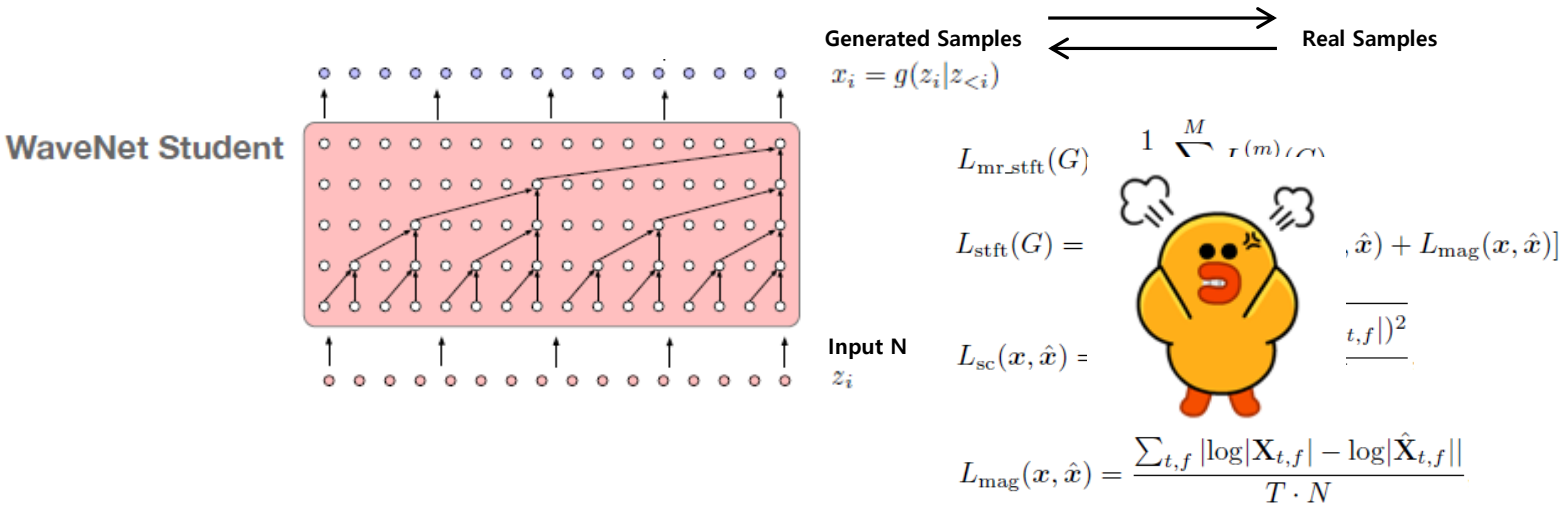
$$L_{\text{mag}}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sum_{t,f}|\log|\mathbf{X}_{t,f}| - \log|\hat{\mathbf{X}}_{t,f}||}{T \cdot N}$$

출처: 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.

Clova

Vocoder: Parallel WaveGAN

1. Teacher-student 기반의 Probability Distillation 과정을 없애고,
2. Adversarial Training 으로 합성음 품질을 높이고,
3. Multi-resolution STFT Loss 로 합성음 품질을 더 높이고,



출처: 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.

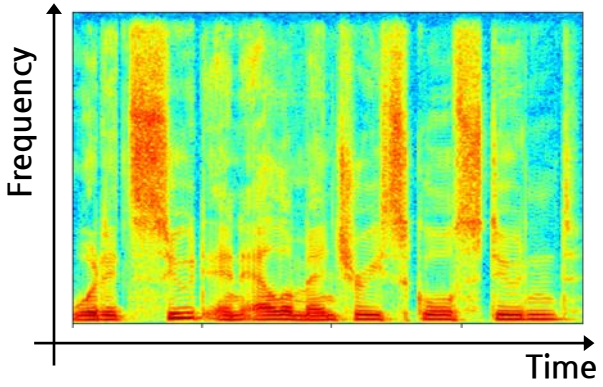
Vocoder: Parallel WaveGAN

1. Teacher-student 기반의 Probability Distillation 과정을 없애고,
2. Adversarial Training 으로 합성음 품질을 높이고,
3. Multi-resolution STFT Loss 로 합성음 품질을 더 높이고,

STFT (short-time Fourier transform)?

짧은 구간의 음성 신호를 주파수 축에서 표현 한 것.

시간 축으로 붙여서 2D 신호로 만든 것이 다음과 같은 Spectrogrm

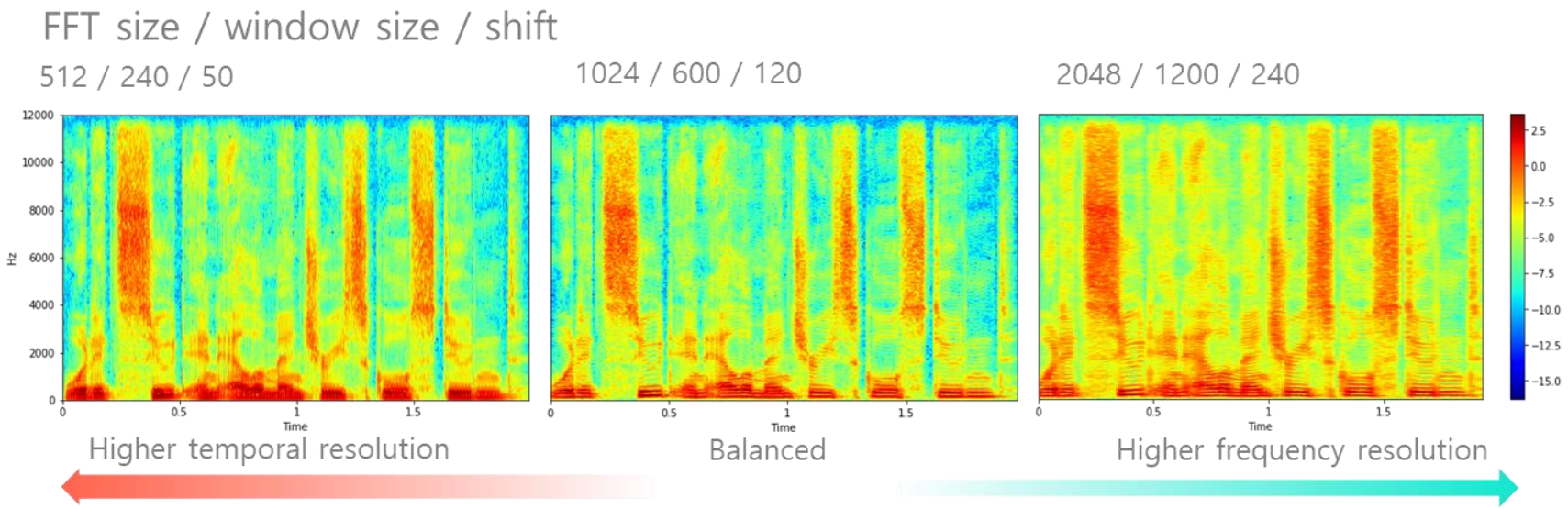


출처: R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

Vocoder: Parallel WaveGAN

1. Teacher-student 기반의 Probability Distillation 과정을 없애고,
2. Adversarial Training 으로 합성음 품질을 높이고,
3. Multi-resolution STFT Loss 로 합성음 품질을 더 높이고,

시간-주파수 축에서 해상도가 다른 여러 개의 Loss 들의 평균이다.



출처: 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.

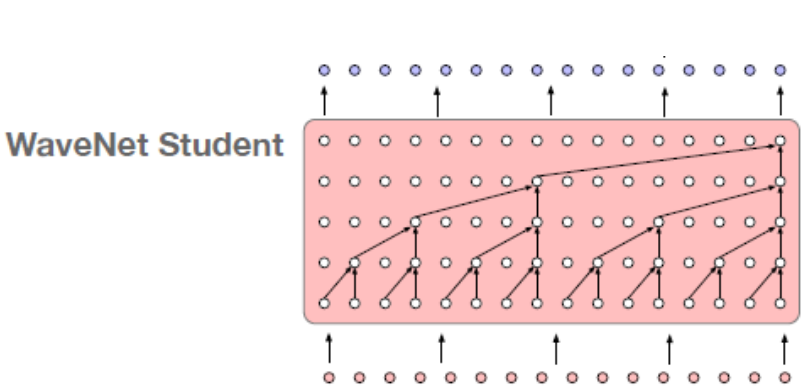
Clova

Vocoder: Parallel WaveGAN

1. Teacher-student 기반의 Probability Distillation 과정을 없애고,
2. Adversarial Training 으로 합성음 품질을 높이고,
3. Multi-resolution STFT Loss 로 합성음 품질을 더 높이고,

시간-주파수 축에서 해상도가 다른 여러 개의 Loss 들의 평균이다.

이때, Loss 는 두가지로 구성되는데



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

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

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

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

출처: 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.

Clova

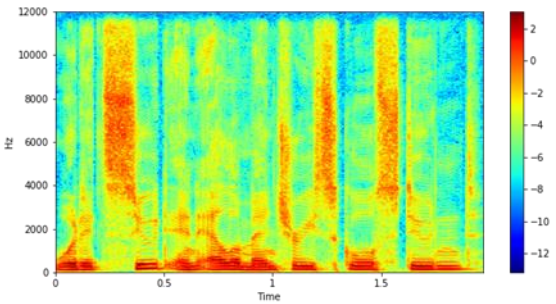
Vocoder: Parallel WaveGAN

1. Teacher-student 기반의 Probability Distillation 과정을 없애고,
2. Adversarial Training 으로 합성음 품질을 높이고,
3. Multi-resolution STFT Loss 로 합성음 품질을 더 높이고,

시간-주파수 축에서 해상도가 다른 여러 개의 Loss 들의 평균이다.

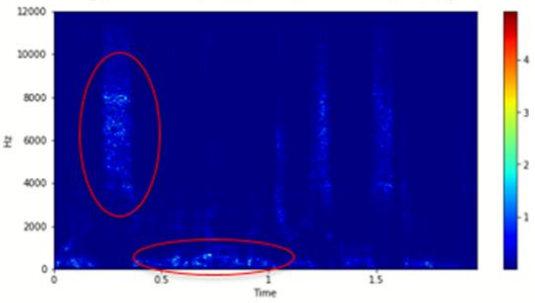
이때, Loss 는 두가지로 구성되는데

하나는 **에너지가 큰 구간**을 잡아내고



Real Spectrogram

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



SC penalizes large amplitude components

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

출처: 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.

Vocoder: Parallel WaveGAN

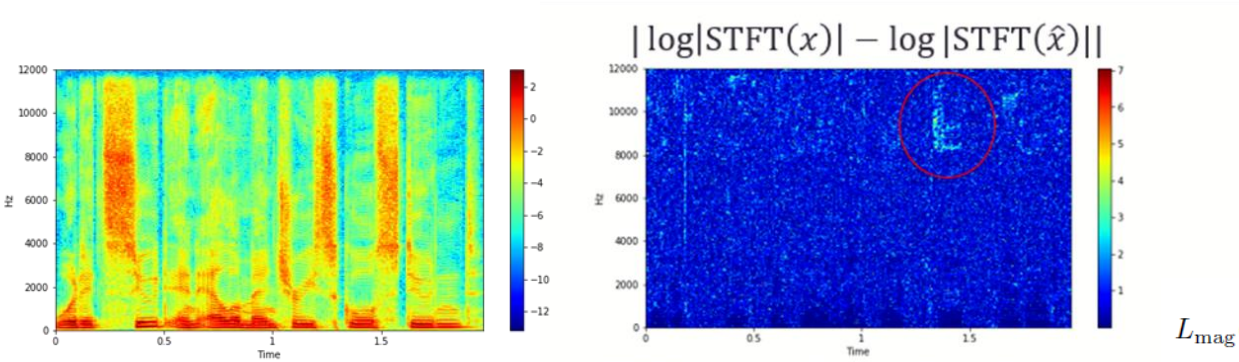
1. Teacher-student 기반의 Probability Distillation 과정을 없애고,
2. Adversarial Training 으로 합성음 품질을 높이고,
3. Multi-resolution STFT Loss 로 합성음 품질을 더 높이고,

시간-주파수 축에서 해상도가 다른 여러 개의 Loss 들의 평균이다.

이때, Loss 는 두가지로 구성되는데

하나는 에너지가 큰 구간을 잡아내고

다른 하나는 **에너지가 작은 구간**을 잡아낸다.



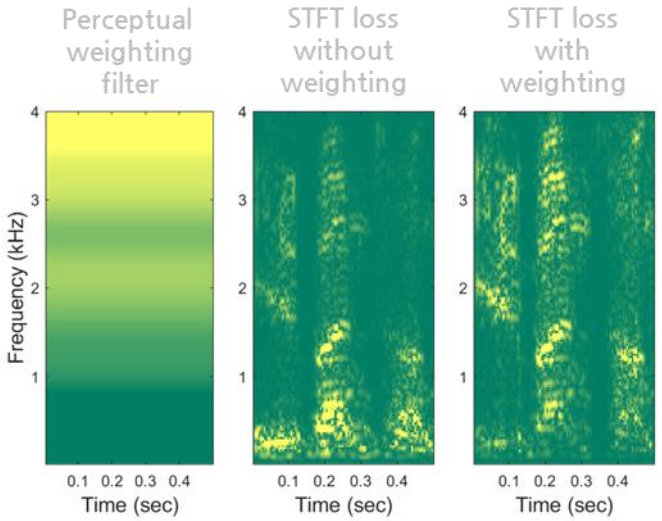
$$L_{\mathrm{mag}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\sum_{t,f} |\log|\mathbf{X}_{t,f}| - \log|\hat{\mathbf{X}}_{t,f}||}{T \cdot N}$$

출처: 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.

Clova

Vocoder: Parallel WaveGAN

1. Teacher-student 기반의 Probability Distillation 과정을 없애고,
2. Adversarial Training 으로 합성음 품질을 높이고,
3. Multi-resolution STFT Loss 로 합성음 품질을 더 높이고,
4. STFT Loss 에 Perceptual Weighting Filter 를 적용해서 한번 더 품질을 높인다.



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

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

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

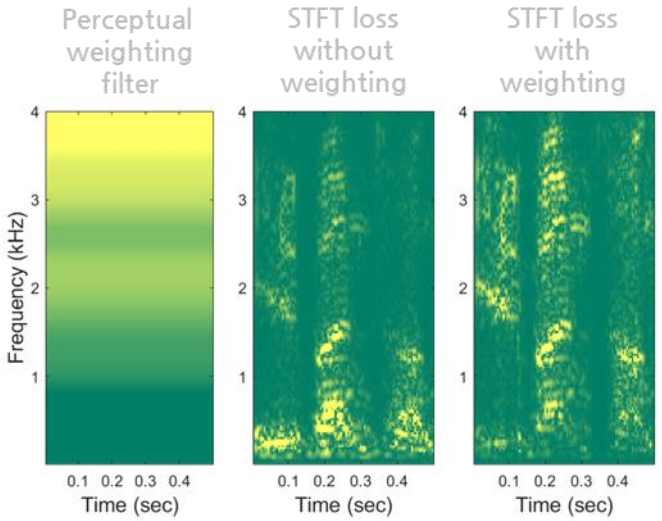
출처: E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” submitted to Proc. SLT, 2021.

Clova

Vocoder: Parallel WaveGAN

1. Teacher-student 기반의 Probability Distillation 과정을 없애고
2. Adversarial Training 으로 합성음 품질을 높이고,
3. Multi-resolution STFT Loss 로 합성음 품질을 더 높이고,
4. STFT Loss 에 Perceptual Weighting Filter 를 적용해서 한번 더 품질을 높인다.

사람에게 청각적으로 더 잘 들리는 **잡음을 제거**하는 역할



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

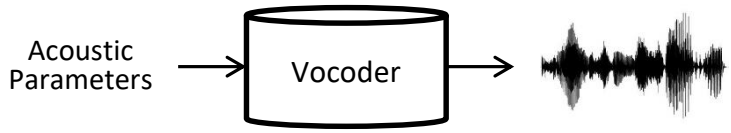
$$L_{\mathrm{mag}}^{w}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sum_{t,f}|\log(\mathbf{W}_{t,f}(\log|\mathbf{X}_{t,f}| - \log|\hat{\mathbf{X}}_{t,f}|)|}{T \cdot N}$$

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

출처: E. Song, et al., “Improved Parallel WaveGAN with perceptually weighted spectrogram loss,” submitted to Proc. SLT, 2021.

Vocoder: Parallel WaveGAN

1. Teacher-student 기반의 Probability Distillation 과정을 없애고,
2. Adversarial Training 으로 합성음 품질을 높이고,
3. Multi-resolution STFT Loss 로 합성음 품질을 더 높이고,
4. STFT Loss 에 Perceptual Weighting Filter 를 적용해서 한번 더 품질을 높인다.



Autoregressive

합성음 품질이 좋지만
생성 속도가 느리다

vs

Non-autoregressive

학습도 쉽고
생성 속도도 빠르고
합성음 품질도 좋다

Performance Evaluations

Evaluation: Database

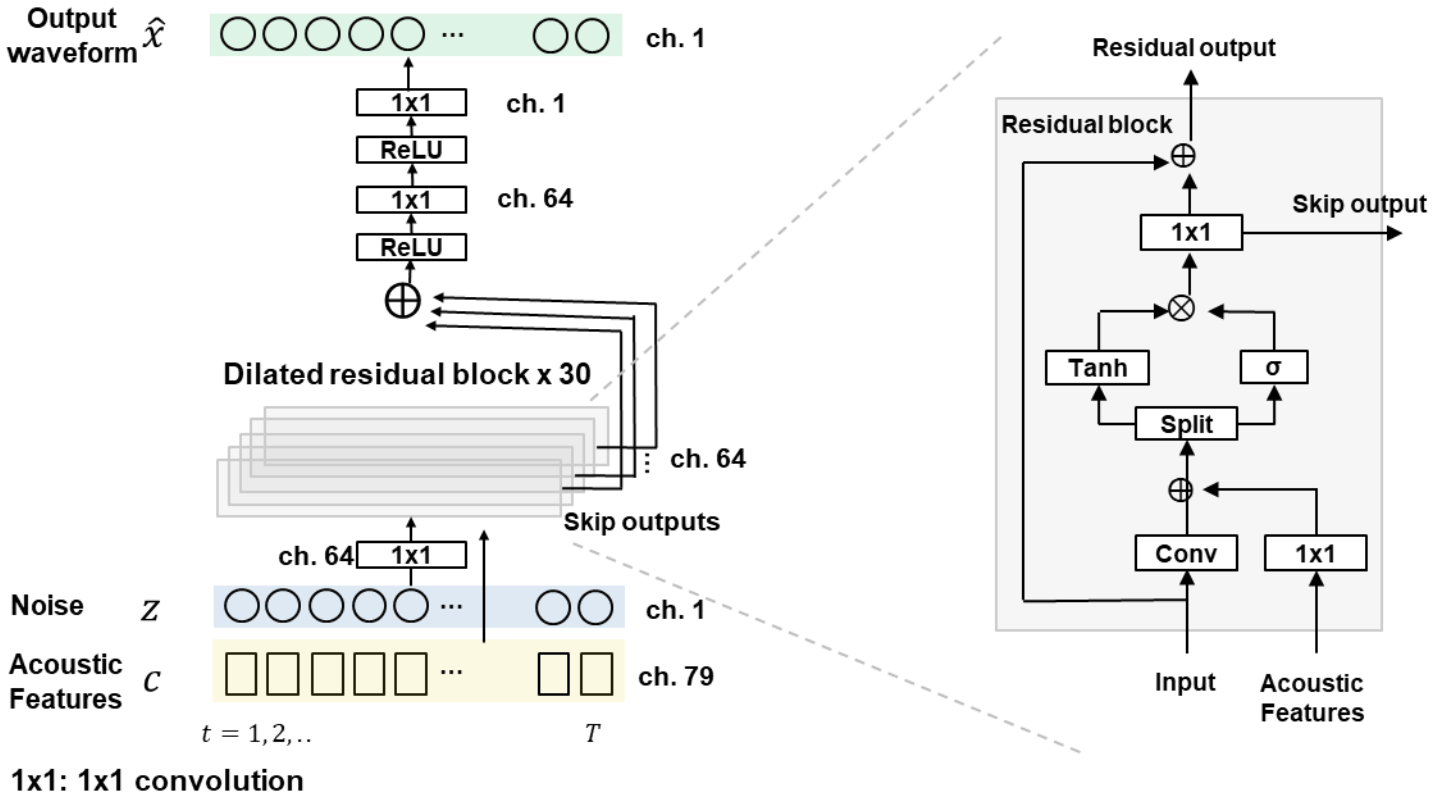
한국 여성 화자 (문장 수)

Training	Validation	Test
5,085	360	180

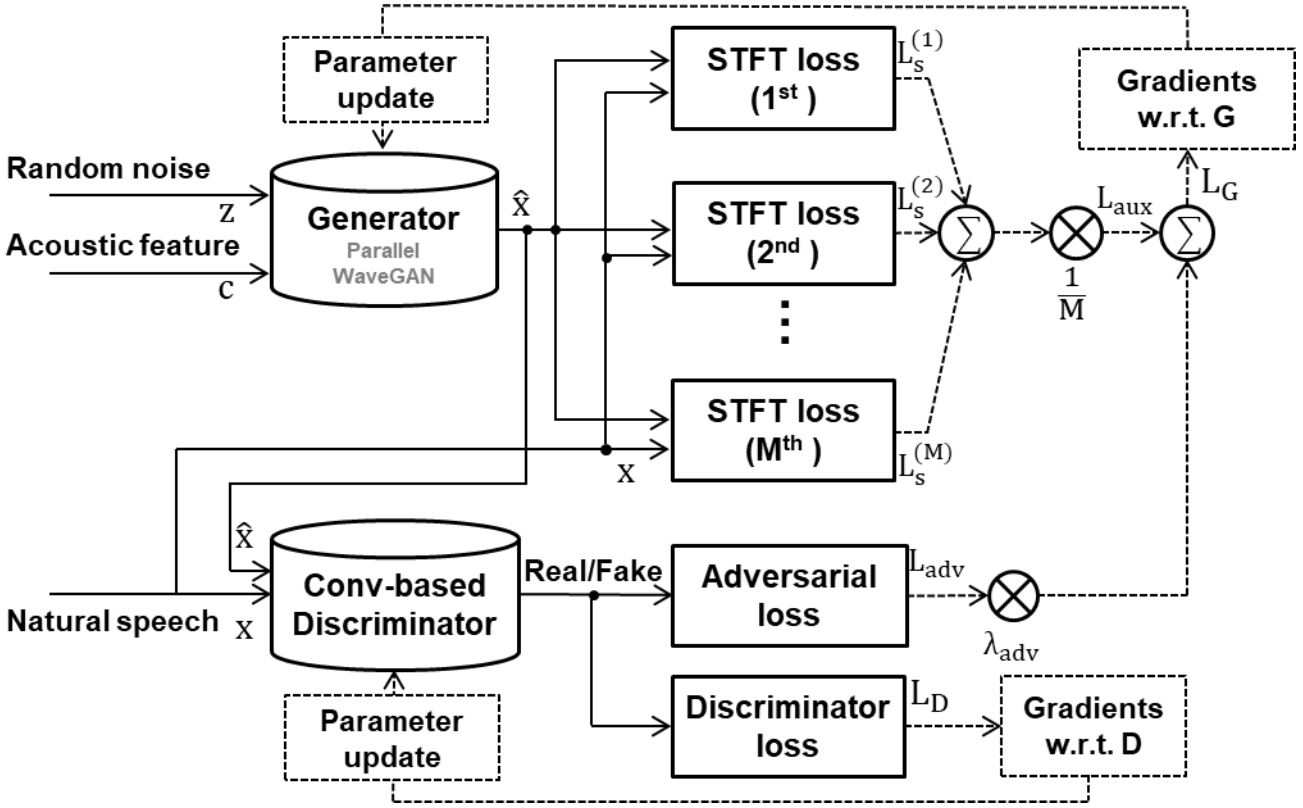
한국 남성 화자 (문장 수)

Training	Validation	Test
5,382	290	140

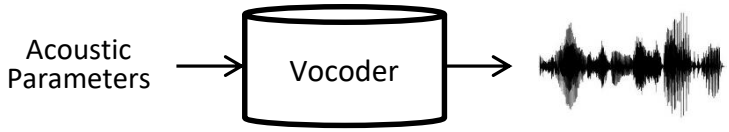
Evaluation: Model Architecture



Evaluation: Training Method



Evaluation: Inference Speed



Autoregressive WaveNet

294.11 RT

VS

Parallel WaveGAN

0.02 RT

약 14,000 배 빠른 생성 속도

RT: 1초 음성을 생성할 때 걸리는 시간 (180+140=320문장)

Evaluation: Perceptual Quality

Autoregressive WaveNet

여성 화자: 3.6 MOS
남성 화자: 3.6 MOS

VS

Parallel WaveGAN

여성 화자: 4.3 MOS
남성 화자: 4.2 MOS

약 14,000 배 빠른 생성 속도
약 18 % 높은 합성음 품질

MOS: 합성음의 자연성을 평가하는 지표 (5.0 만점)

Evaluation: TTS Demo

Autoregressive WaveNet

여성 화자:
남성 화자:

VS

Parallel WaveGAN

여성 화자:
남성 화자:

약 14,000 배 빠른 생성 속도
약 18 % 높은 합성음 품질

Evaluation: TTS Demo

Autoregressive WaveNet

여성 화자:
남성 화자:

VS

Parallel WaveGAN

여성 화자:
남성 화자:

약 14,000 배 빠른 생성 속도
약 18 % 높은 합성음 품질

Summary

본 발표에서는 TTS 엔진의 핵심 요소인 **Vocoder** 기술을 정리하고,
빠르고 안정적인 **Parallel WaveGAN** 모델을 소개하였습니다.

Clova

Goal

빠르고 안정적인 NON-autoregressive Vocoder 만들기

How to ?

Teacher-student 기반의 Probability Distillation 과정을 없애고,
Adversarial Training 으로 합성음 품질을 높이고,
Multi-resolution STFT Loss 로 합성음 품질을 더 높이고,
STFT Loss 에 Perceptual Weighting Filter 를 적용해서 한번 더 품질을 높인다.

Parallel WaveGAN's performance

14,000 배 빠른 생성 속도 + 18 % 높은 합성음 품질

본 발표에서는 TTS 엔진의 핵심 요소인 **Vocoder** 기술을 정리하고,
빠르고 안정적인 **Parallel WaveGAN** 모델을 소개하였습니다.

Clova

Next ?

서비스 적용을 위한 안정성 높이기

TTS samples

한국어

일본어

기술 개발 기여자 (가나다순)

HDTS/권오성
HDTS/김진섭
HDTS/송은우
HDTS/송찬호
HDTS/황민제
HDTS/山本龍一
Voice&Avatar/김재민

Q / A