ENGiNEERiNG DAY

Harmonic-plus-Noise Parallel WaveGAN

빠르고, 품질 좋은 WaveNet 음성 합성기 만들기

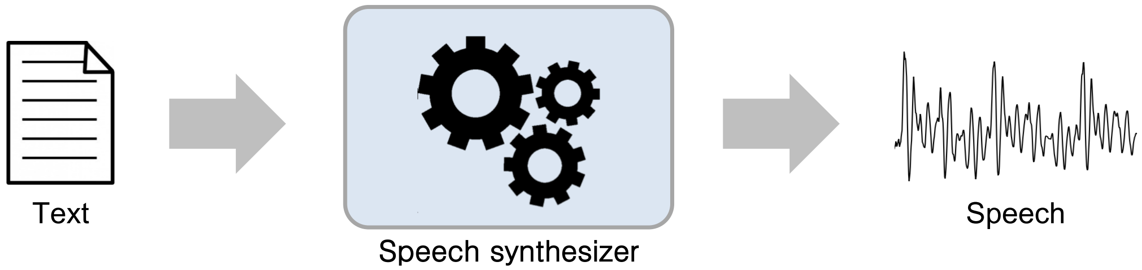
황민제 / HDTS

NAVER CLOVA

CONTENTS

Introduction

Text-to-Speech (TTS) technology

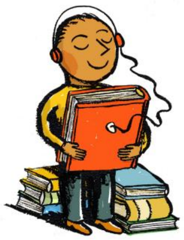


- The system synthesizing speech waveform from given input text

Application area

Navigation

AI speaker



Audiobook

Ai Call

Speech translation

INTRODUCTION

TTS system overview

[End-to-end TTS system]

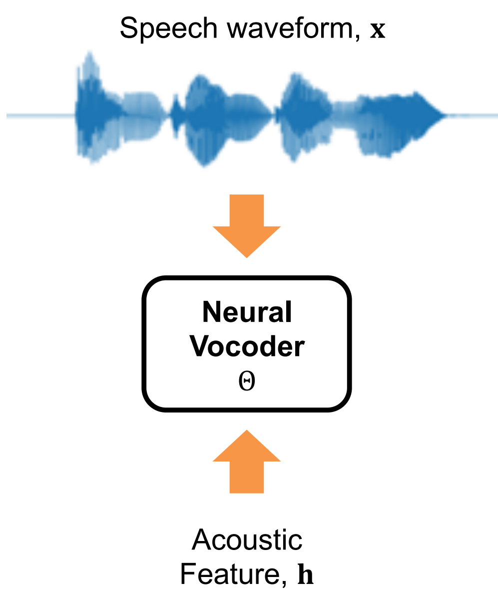
- Acoustic model
  - Generate speech's acoustic feature from input text
  - Acoustic features?
    - Mel-spectrum / pitch / energy / voicing information, ...
  - Famous model [1, 2]
    - Tacotron / FastSpeech, ...
- Neural vocoder
  - Synthesize speech waveform from generated acoustic features
  - Famous model [3]
    - WaveNet..

[1] Shen et. al., “Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions,” in *CoRR*, 2017.
[2] Ren at al., "FastSpeech: Fast, Robust and Controllable Text to Speech," in *NeurIPS*, 2019
[3] Aaron et al., "WaveNet: A Generative Model for Raw Audio," in *Arxiv*, 2016

Neural vocoder

[Training phase]

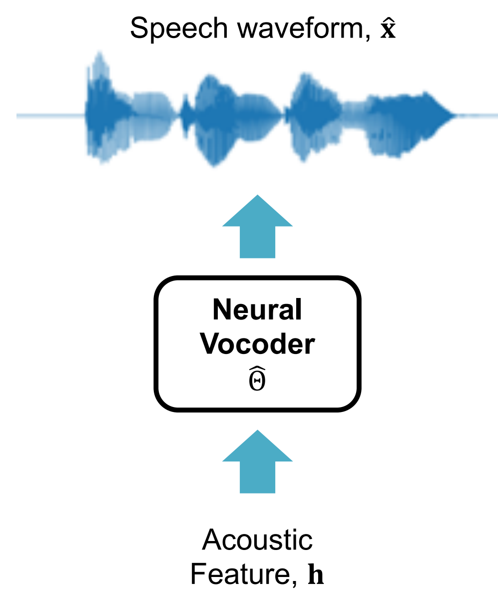
$$\hat{\Theta} = \arg\max_{\Theta} p(\mathbf{x} \mid \mathbf{h}, \Theta)$$



Optimize network parameters to maximize the likelihood of speech waveform

[Inference phase]

$$\hat{\mathbf{x}} \sim p(\mathbf{x} \mid \mathbf{h}, \hat{\Theta})$$



Sample speech waveform from estimated speech likelihood

NEURAL VOCODER

[Training phase]

$$\hat{\Theta} = \arg\max_{\Theta} p(\mathbf{x} \mid \mathbf{h}, \Theta)$$

Speech waveform, $\mathbf{x}$

Neural Vocoder
Θ

Acoustic Feature, $\mathbf{h}$

Optimize network parameters to maximize the likelihood of speech waveform

[Inference phase]

$$\hat{\mathbf{x}} \sim p(\mathbf{x} \mid \mathbf{h}, \hat{\Theta})$$

Speech waveform, $\hat{\mathbf{x}}$

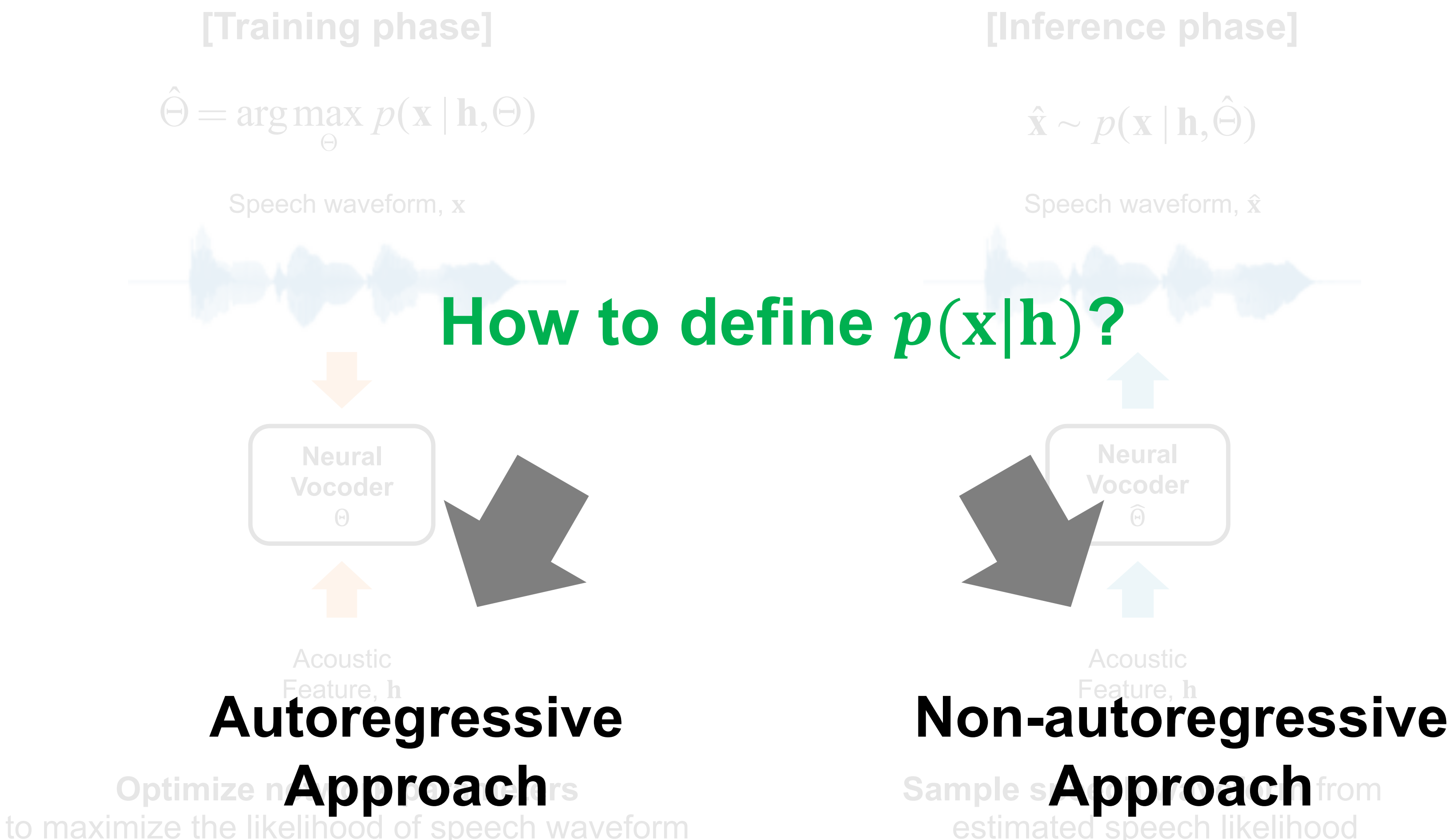
Neural Vocoder
$\hat{\Theta}$

Acoustic Feature, $\mathbf{h}$

Sample speech waveform from estimated speech likelihood

How to define $p(\mathrm{x}|\mathrm{h})$?

Neural vocoder



How to define $p(\mathbf{x}|\mathbf{h})$?

Autoregressive Approach

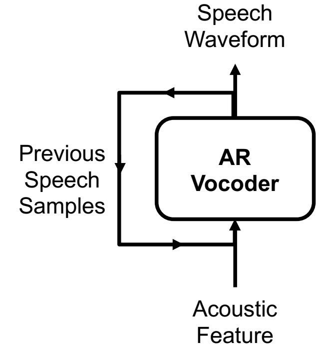
Non-autoregressive Approach

WaveNet

Autoregressive (AR) modeling for audio waveform [3]

$$p(\mathbf{x}\,|\,\mathbf{h}) = \prod_{n=0}^{T-1} p(x_n\,|\,\mathbf{x}_{<n}, \mathbf{h})$$

- Input
  - Acoustic features
  - Previously generated waveform samples



[Concept of AR vocoder]

Key structure

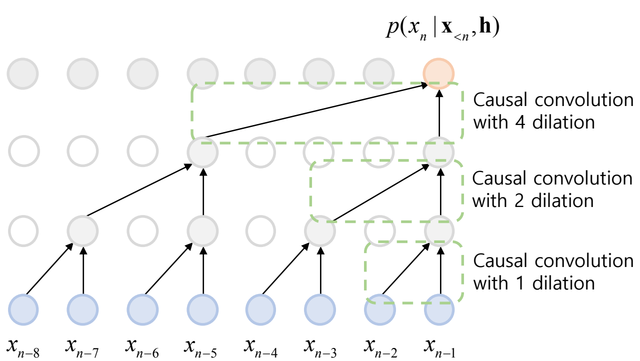
- Stack of dilated causal convolution
  - Result in exponentially growing receptive field
  - Effectively capture speech's long-term dependency property

Advantage

- Provide significantly better synthesis quality than conventional vocoders

Problem

- Very slow generation speed
  - 300 real-time factor (RTF)
    = require 300 sec. for synthesizing 1 sec. of speech



[WaveNet]

[3] Aaron et al., "WaveNet: A Generative Model for Raw Audio," in *Arxiv*, 2016

PARALLEL WAVEGAN (PWG)

WaveNet for non-AR neural vocoder [4]

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{n=0}^{T-1} p(x_n \mid \mathbf{h})$$

- Input
  - Acoustic features
  - Gaussian noise

Key structure

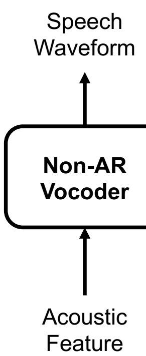
- (1) Non-causal WaveNet + (2) Adversarial training
  - (1) ***Enable fast generation***
  - (2) ***Prevent quality degradation caused by non-AR modeling***

Advantage

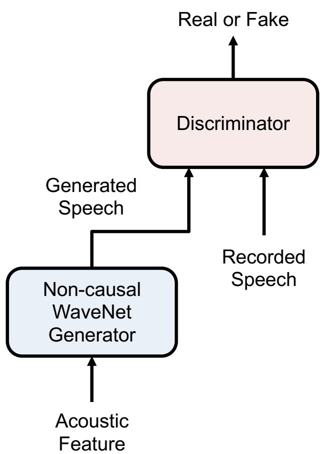
- Very fast synthesis speed
  - 0.02 RTF = 15,000 times faster than WaveNet

Problem

- Unstable, and low quality of synthesized speech



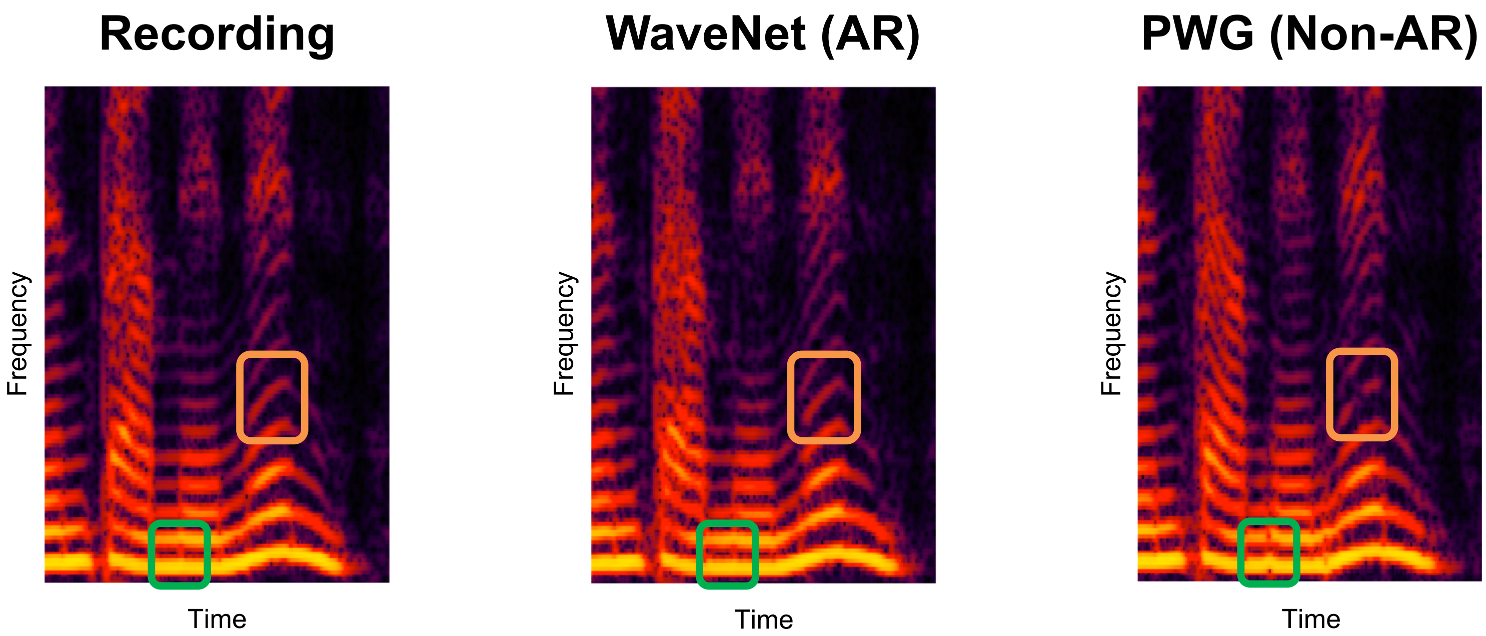
[Concept of non-AR vocoder]



[PWG]

[4] R. Yamamoto et. al., "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," *in Proc. ICASSP*, 2020.

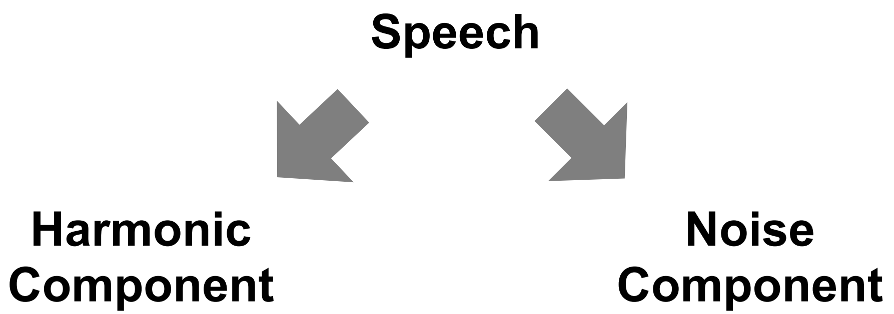
SPECTROGRAM EXAMPLE



HARMONIC-PLUS-NOISE PWG (HN-PWG)

Adopt harmonic-plus-noise (HN) model to the PWG's generator

- HN model [5]**?**
  - **speech = *harmonic component* + *noise component***
    = Periodic, deterministic *= Aperiodic, stochastic*

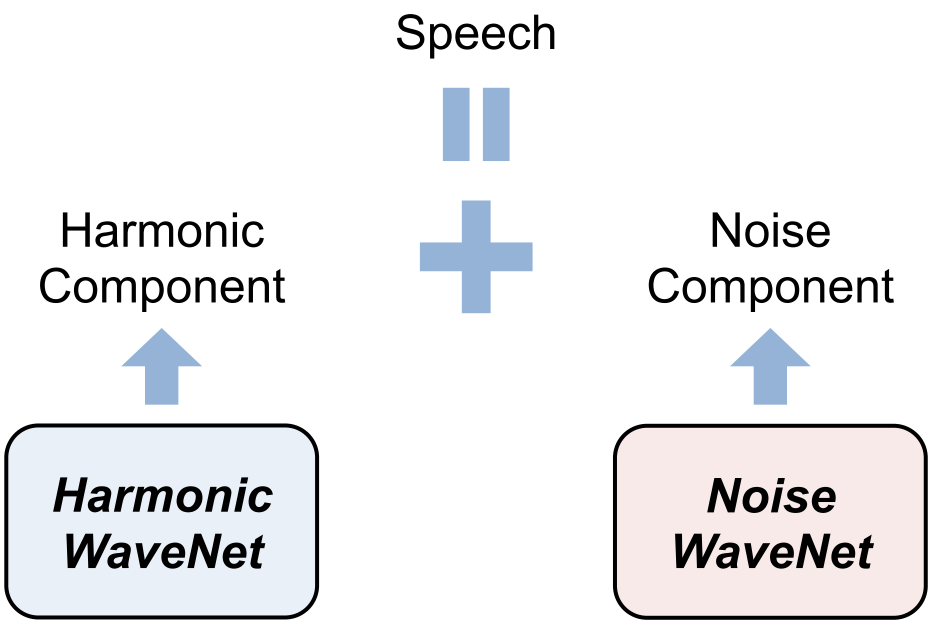


[5] Y. Stylianou, "Modeling speech based on harmonic plus noise models," in Nonlinear Speech Modeling and Applications. Springer Berlin Heidelberg, 2005.

HARMONIC-PLUS-NOISE PWG (HN-PWG)

Adopt harmonic-plus-noise (HN) model [5] to the PWG's generator

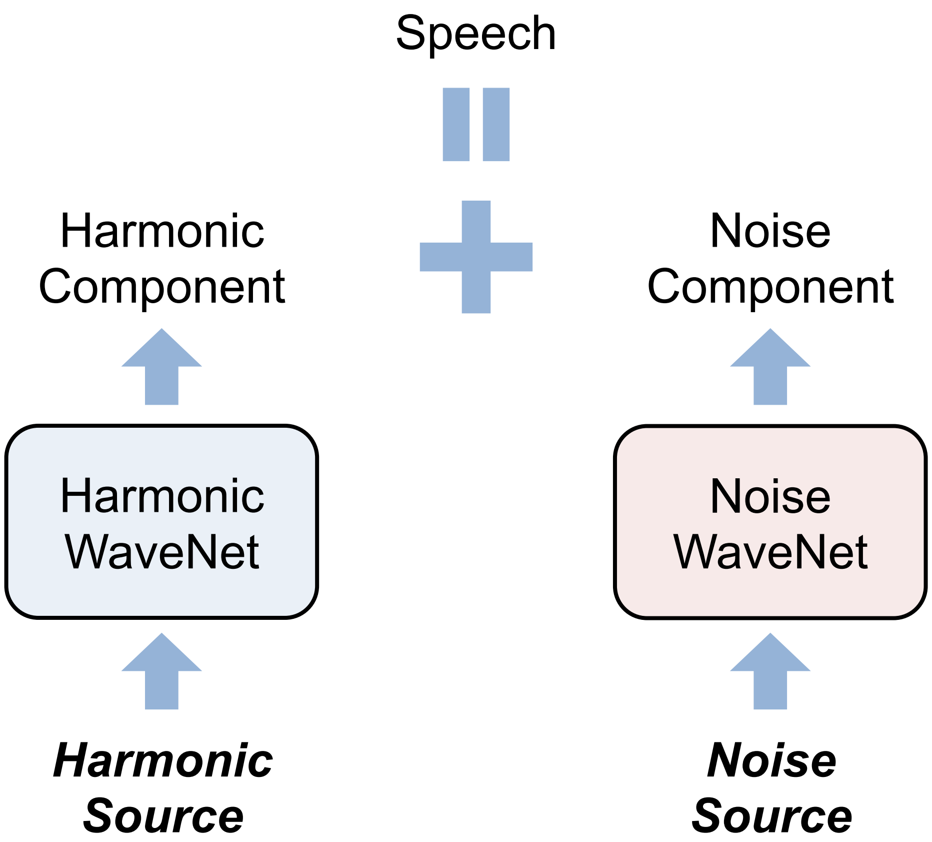
- Split WaveNet generator to two sub-WaveNet generators
  1. Harmonic WaveNet (H-WaveNet) → Generate harmonic component
  2. Noise WaveNet (N-WaveNet) → Generate noise component



HARMONIC-PLUS-NOISE PWG (HN-PWG)

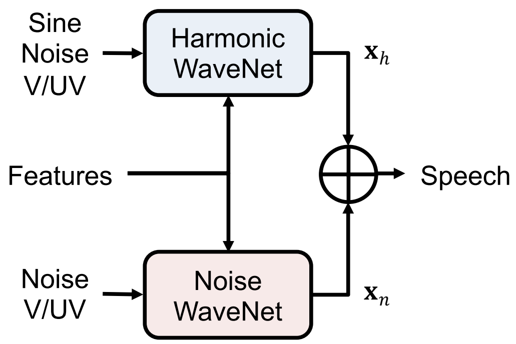
Adopt harmonic-plus-noise (HN) model [5] to the PWG's generator

- Method to impose harmonic & noise characteristics
  - Feeding harmonic- and noise-like sources to their WaveNets, respectively



HARMONIC-PLUS-NOISE PWG (HN-PWG)

Concept of HN-PWG



[HN-PWG]

Source signal designs

1. H-WaveNet
   - Give harmonic (=periodic) characteristic by using sinusoidal source signal

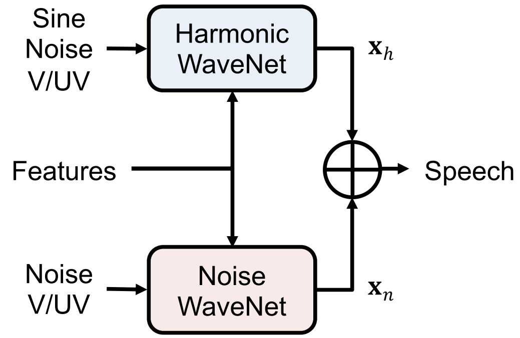
$$s[t] = \sin\left(\sum_{k=1}^{t} 2\pi \frac{f_k}{F_s} + \phi\right)$$

     - Design source signal to have instantaneous frequency of pitch contour

2. N-WaveNet
   - Give noise (=aperiodic) characteristic by using Gaussian noise source signal

HARMONIC-PLUS-NOISE PWG (HN-PWG)

Concept of HN-PWG



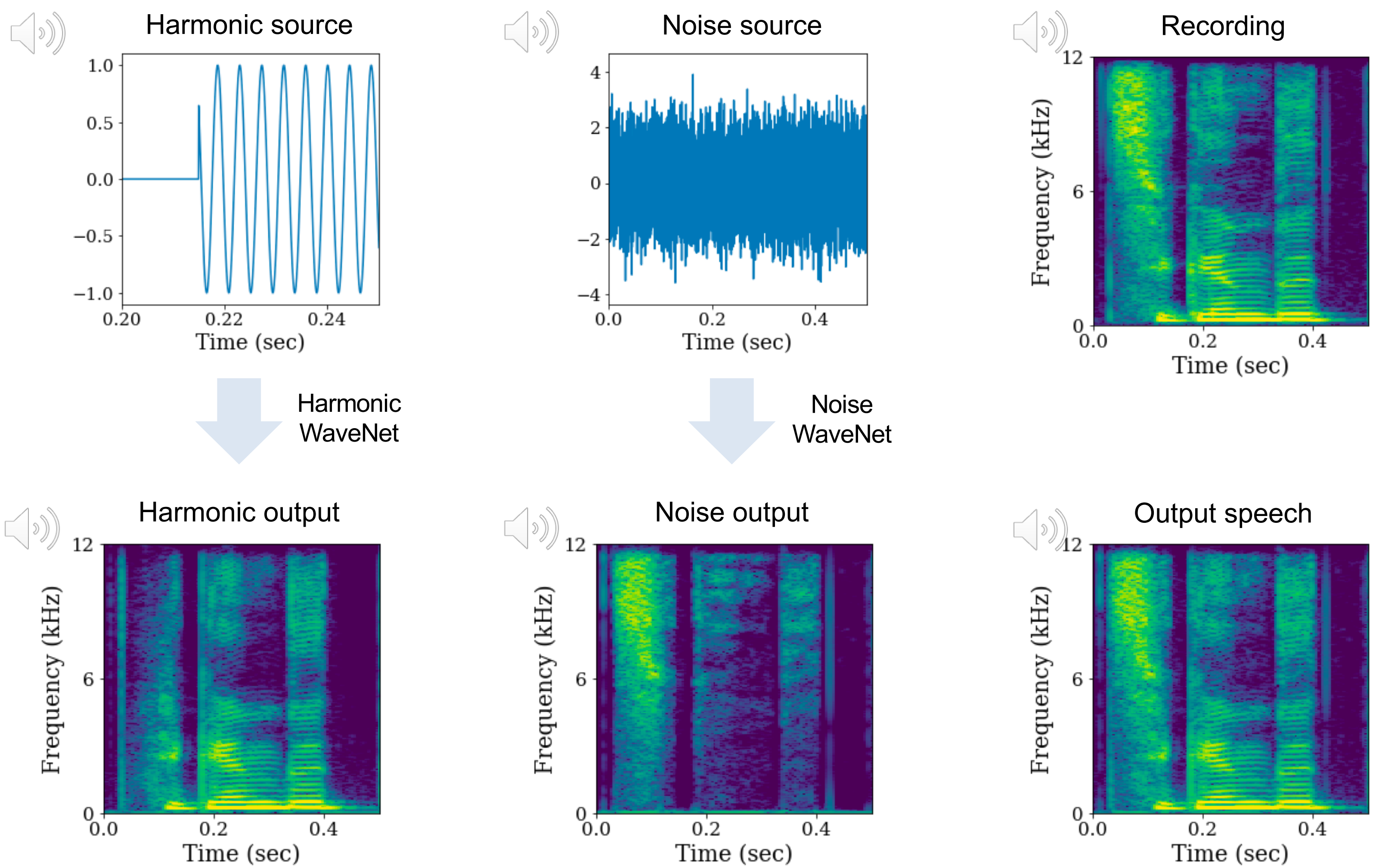
[HN-PWG]

Additional sources

1. H-WaveNet
   - Sequence of voicing flag (V/UV)
     - Enable each WaveNet to be effectively aware of voicing state
   - Gaussian noise
     - Empirically improve synthesis quality
2. N-WaveNet
   - Sequence of V/UV

Harmonic-plus-Noise PWG (HN-PWG)

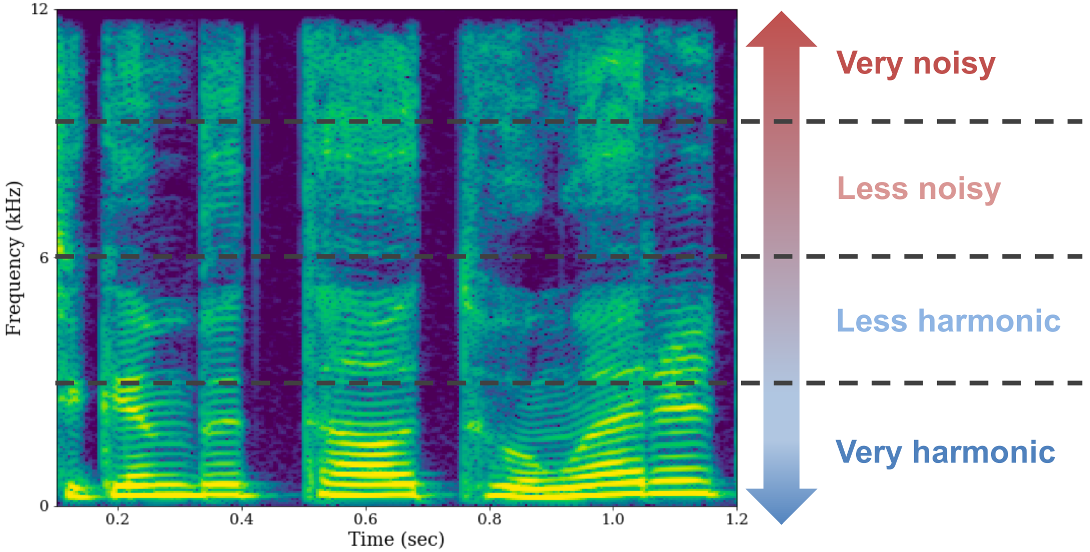
Speech sample



Multi-band HN-PWG

Motivation to further improve HN-PWG's performance

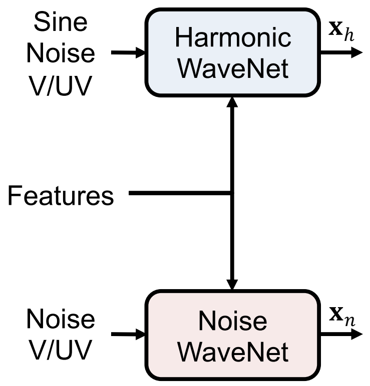
- Consider harmonic-noise property of speech signal
  - Low frequency band
    - Harmonic characteristic > Noise characteristic
  - High frequency band
    - Harmonic characteristic < Noise characteristic



Introduce this harmonic-noise property to the HN-PWG

MULTI-BAND HN-PWG

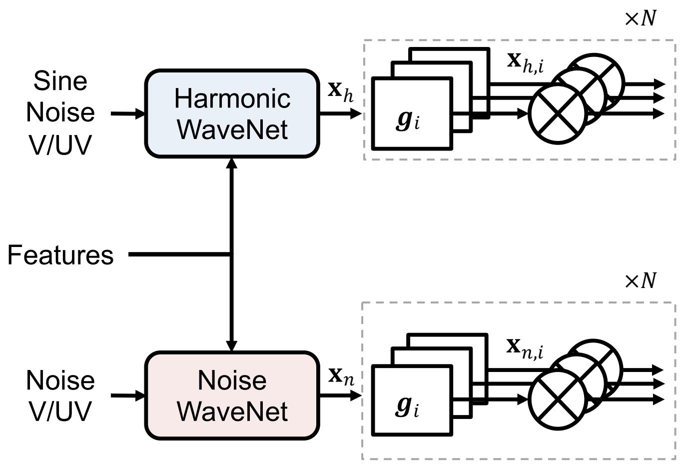
Multi-band HN-PWG



Step 1.
Generate harmonic component x_h and noise component x_n by using H- and N-WaveNets

Multi-band HN-PWG

Multi-band HN-PWG



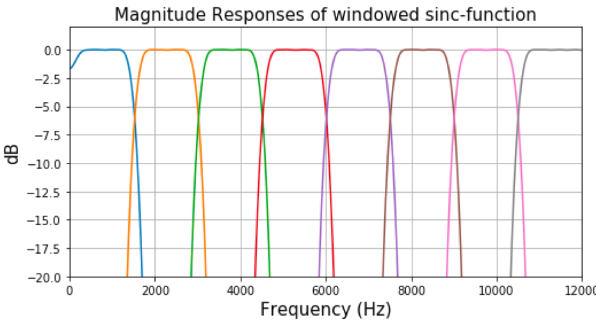
Step 2.
Decompose generated harmonic-noise components into **their subband signals** by using **windowed sinc function-based band-pass filters** (BPF; $\mathbf{g}_i$)

$$\mathbf{x}_{h,i} = \mathbf{x}_h \circledast \hat{\mathbf{g}}_i$$

$$\mathbf{x}_{n,i} = \mathbf{x}_n \circledast \hat{\mathbf{g}}_i$$

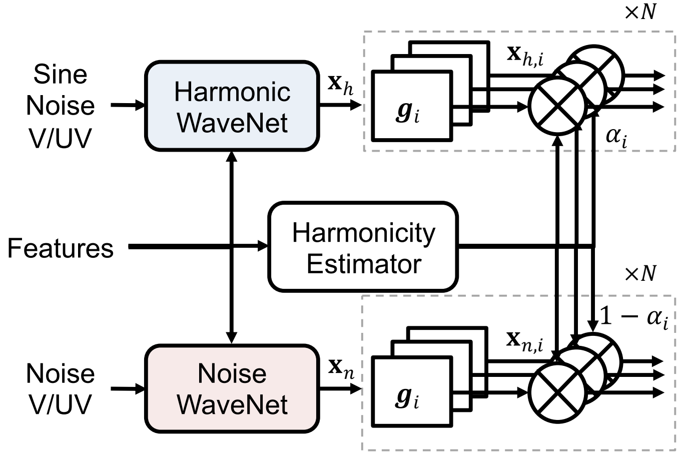
$$\text{where} \quad g_i[k] = 2f_{i+1}\operatorname{sinc}(2\pi f_{i+1}k) - 2f_i\operatorname{sinc}(2\pi f_i k),$$

$$\hat{g}_i[k] = g_i[k] \cdot w_{hamm}[k]$$



Multi-band HN-PWG

Multi-band HN-PWG



Step 3.
Estimate subband harmonicity from acoustic features

$$\{\alpha_i\} = sigmoid(CNN(\mathbf{h}))$$

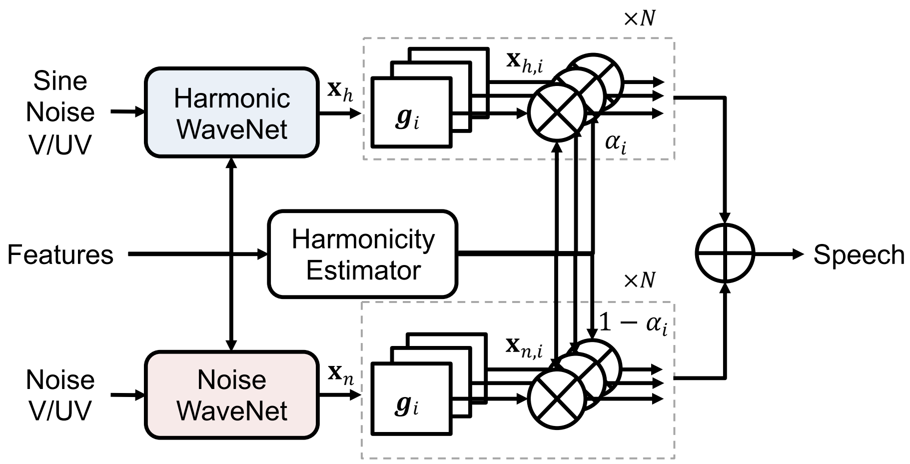
Then, adjust gain of subband signals weighted by subband harmonicity

$$\hat{\mathbf{x}}_{h,i} = \alpha_i \cdot \mathbf{x}_{h,i}$$

$$\hat{\mathbf{x}}_{n,i} = (1-\alpha_i) \cdot \mathbf{x}_{h,i}$$

MULTI-BAND HN-PWG

Multi-band HN-PWG



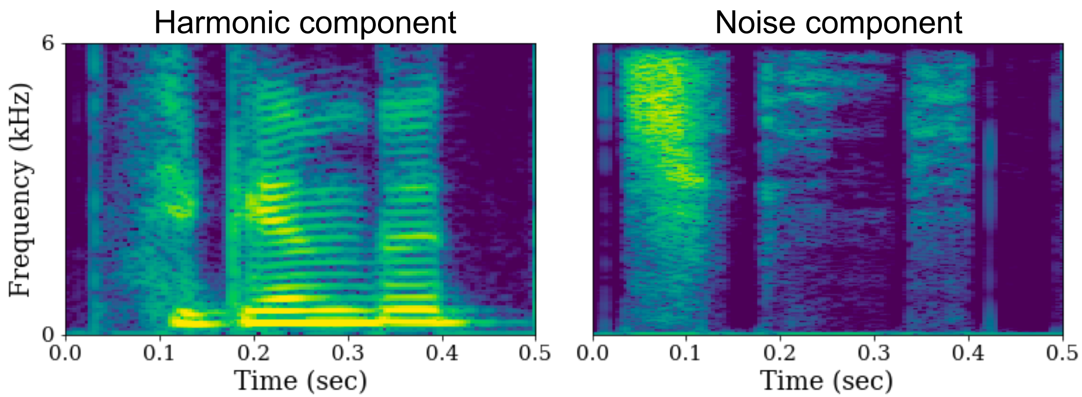
Step 4.
Sum all of subband signals

$$\mathbf{x} = \sum_{i=0}^{N-1} [\hat{\mathbf{x}}_{h,i} + \hat{\mathbf{x}}_{n,i}]$$

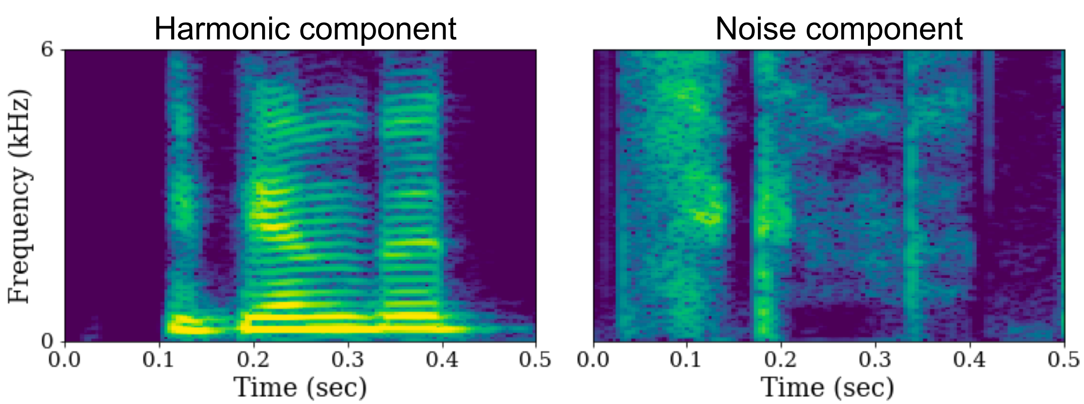
Multi-band HN-PWG

Spectrogram comparison with HN-PWG

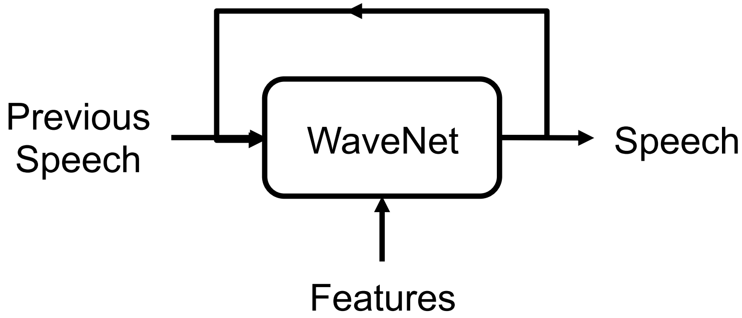
- HN-PWG



- Multi-band HN-PWG



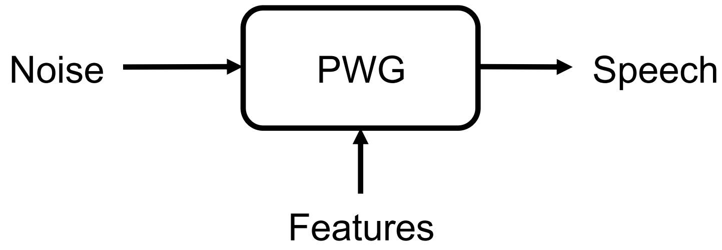
SUMMARY



[WaveNet]

AR model for speech waveform
☺ Good quality
☹ Slow generation speed

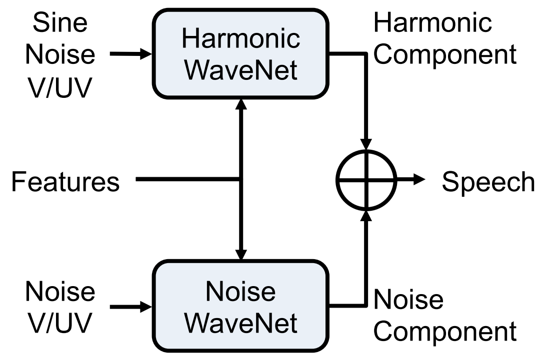
SUMMARY



[PWG]

Non-AR WaveNet + GAN framework
☺ Fast generation speed
☹ Unsatisfactory synthesis quality

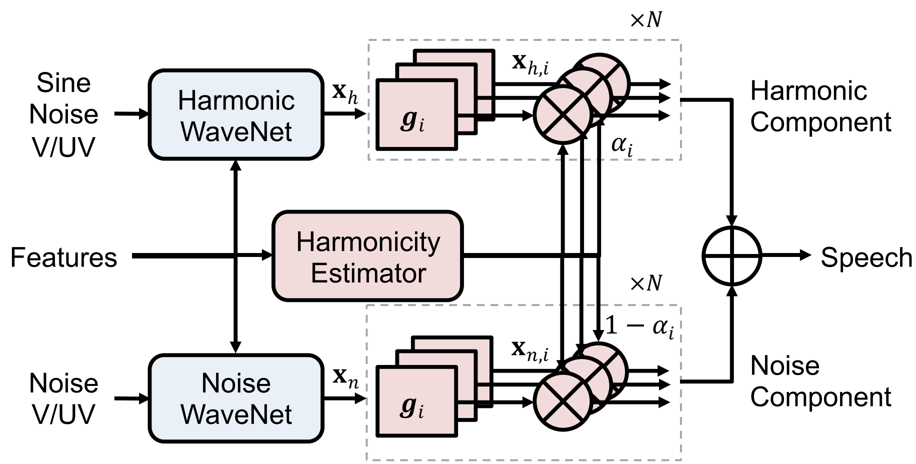
SUMMARY



[HN-PWG]

Adopt HNM to the PWG generator

SUMMARY



[Multi-band HN-PWG]

Adopt Multi-band HNM to the PWG generator

EXPERIMENTS

Database

- Korean female speaker
- Sampling rate / quantization
  - 24-kHz / 16-bit
- Acoustic features
  - Improved time-frequency trajectory excitation (ITFTE) vocoder [6]

Neural vocoders

Model	Use of HN model	Input signals for H-WaveNet	Type of HN model
WaveNet	-	-	-
PWG	-	-	-
HN-PWG w/o noise	Yes	Sine + V/UV	Full-band
HN-PWG	Yes	Sine + noise + V/UV	Full-band
Multi-band HN-PWG	Yes	Sine + noise + V/UV	Multi-band

[6] E. Song, et. al., "Effective spectral and excitation modeling techniques for LSTM-RNN-based speech synthesis systems," in IEEE/ACM Trans. ASLP, 2017.

EXPERIMENTS

Evaluation metrics

- Model size
  - Number of parameters consisting neural vocoder
- Inference speed
  - Measure real-time factor (RTF) on single V100 GPU
- Mean opinion score (MOS) listening test
  - Score the subjective quality of speech (from 1.0 to 5.0)
  - Analysis / synthesis scenario
    - Use ground-truth acoustic features
  - TTS scenario
    - Use generated acoustic features from TTS model

[Scoring criteria for MOS test]

Score	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

EXPERIMENTS

Results

Model	Model size ↓ (M)	Inference speed ↓ (RTF)	MOS ↑	
			Analysis / synthesis scenario	TTS scenario
WaveNet	3.81	294.12	4.22	4.03
PWG	0.94	0.02	3.46	3.56
HN-PWG w/o noise	0.94	0.02	4.02	2.60
HN-PWG	0.94	0.02	4.18	4.01
Multi-band HN-PWG	0.99	0.02	4.29	4.03
Recordings	-	-	4.41	

EXPERIMENTS

Results

Model	Model size ↓ (M)	Inference speed ↓ (RTF)	MOS ↑	
			Analysis / synthesis scenario	TTS scenario
WaveNet	**3.81**	**294.12**	4.22	4.03
PWG	**0.94**	**0.02**	3.46	3.56
HN-PWG w/o noise	**0.94**	**0.02**	4.02	2.60
HN-PWG	**0.94**	**0.02**	4.18	4.01
Multi-band HN-PWG	**0.99**	**0.02**	4.29	4.03
Recordings	-	-	4.41	

1. **Non-AR models provided significantly faster synthesis speed and smaller network size than AR-WaveNet.**

EXPERIMENTS

Results

Model	Model size ↓ (M)	Inference speed ↓ (RTF)	MOS ↑	
			Analysis / synthesis scenario	TTS scenario
WaveNet	3.81	294.12	4.22	4.03
PWG	**0.94**	**0.02**	3.46	3.56
HN-PWG w/o noise	**0.94**	**0.02**	4.02	2.60
HN-PWG	**0.94**	**0.02**	4.18	4.01
Multi-band HN-PWG	**0.99**	**0.02**	4.29	4.03
Recordings	-	-	4.41	

1. Non-AR models provided significantly faster synthesis speed and smaller network size than AR-WaveNet.
2. **Use of HN model didn't affect the model size and inference speed.**

EXPERIMENTS

Results

Model	Model size ↓ (M)	Inference speed ↓ (RTF)	MOS ↑	
			Analysis / synthesis scenario	TTS scenario
WaveNet	3.81	294.12	**4.22**	**4.03**
PWG	0.94	0.02	**3.46**	**3.56**
HN-PWG w/o noise	0.94	0.02	4.02	2.60
HN-PWG	0.94	0.02	4.18	4.01
Multi-band HN-PWG	0.99	0.02	4.29	4.03
Recordings	-	-	4.41	

1. Non-AR models provided significantly faster synthesis speed and smaller network size than AR-WaveNet.
2. Use of HN model didn't affect the model size and inference speed.
3. **Conventional PWG showed worse quality than WaveNet.**

EXPERIMENTS

Results

Model	Model size ↓ (M)	Inference speed ↓ (RTF)	MOS ↑	
			Analysis / synthesis scenario	TTS scenario
WaveNet	3.81	294.12	4.22	4.03
PWG	0.94	0.02	**3.46**	**3.56**
HN-PWG w/o noise	0.94	0.02	4.02	2.60
HN-PWG	0.94	0.02	**4.18**	**4.01**
Multi-band HN-PWG	0.99	0.02	**4.29**	**4.03**
Recordings	-	-	4.41	

1. Non-AR models provided significantly faster synthesis speed and smaller network size than AR-WaveNet.
2. Use of HN model didn't affect the model size and inference speed.
3. Conventional PWG showed worse quality than WaveNet.
4. **However, its quality was significantly improved by adopting HN model.**

Experiments

Results

Model	Model size ↓ (M)	Inference speed ↓ (RTF)	MOS ↑	
			Analysis / synthesis scenario	TTS scenario
WaveNet	3.81	294.12	4.22	4.03
PWG	0.94	0.02	3.46	3.56
HN-PWG w/o noise	0.94	0.02	**4.02**	**2.60**
HN-PWG	0.94	0.02	4.18	4.01
Multi-band HN-PWG	0.99	0.02	4.29	4.03
Recordings	-	-	4.41	

1. Non-AR models provided significantly faster synthesis speed and smaller network size than AR-WaveNet.
2. Use of HN model didn't affect the model size and inference speed.
3. Conventional PWG showed worse quality than WaveNet.
4. However, its quality was significantly improved by adopting HN model.
5. **In TTS scenario, the quality of HN-PWG became severely degraded when the noise source is not us ed for H-WaveNet.**

EXPERIMENTS

Results

Model	Model size ↓ (M)	Inference speed ↓ (RTF)	MOS ↑	
			Analysis / synthesis scenario	TTS scenario
WaveNet	3.81	294.12	**4.22**	**4.03**
PWG	0.94	0.02	3.46	3.56
HN-PWG w/o noise	0.94	0.02	4.02	2.60
HN-PWG	0.94	0.02	4.18	4.01
Multi-band HN-PWG	0.99	0.02	**4.29**	**4.03**
Recordings	-	-	4.41	

1. Non-AR models provided significantly faster synthesis speed and smaller network size than AR-WaveNet.
2. Use of HN model didn't affect the model size and inference speed.
3. Conventional PWG showed worse quality than WaveNet.
4. However, its quality was significantly improved by adopting HN model.
5. In TTS scenario, the quality of HN-PWG became severely degraded when the noise source is not used for H-WaveNet.
6. **Use of multi-band HN model improved quality of HN-PWG, and even better than AR WaveNet.**

SUMMARY & CONCLUSION

Proposed Harmonic-plus-Noise (HN) Parallel WaveGAN (PWG) vocoder

Problems of conventional vocoders

- WaveNet: Good quality, but slow speed
- PWG: Fast speed, but unsatisfactory quality

Proposed HN-PWG = Fast and high-quality neural vocoder

- HN-PWG
  - Apply HN model to PWG's generator architecture
- Multi-band HN-PWG
  - Apply multi-band HN model to HN-PWG

Experimental results

- Provided significantly better quality than conventional vocoders while maintaining fast synthe sis speed

SUMMARY & CONCLUSION

Will be published at the conference of Interspeech 2021

More questions

- min-jae.hwang@navercorp.com

References

[1] Shen et. al., “Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions,” in *CoRR*, 2017.

[2] Y. Ren et. al., "FastSpeech 2: Fast and High-Quality End-toEnd Text to Speech," *in Arxiv*, 2020.

[3] Aaron et al., "WaveNet: A Generative Model for Raw Audio," in *Arxiv*, 2016

[4] R. Yamamoto et. al., "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," *in Proc. ICASSP,* 2020.

[5] Y. Stylianou, “Modeling speech based on harmonic plus noise models,” in Nonlinear Speech Modeling and Applications. Spring er Berlin Heidelberg, 2005.

[6] E. Song, et. al., “Effective spectral and excitation modeling techniques for LSTM-RNN-based speech synthesis systems,” in IEE E/ACM Trans. ASLP, 2017.

Thank you!