TTS-driven Data Augmentation

가짜 목소리 DB로 고품질 음성 합성기를 만들어보자!

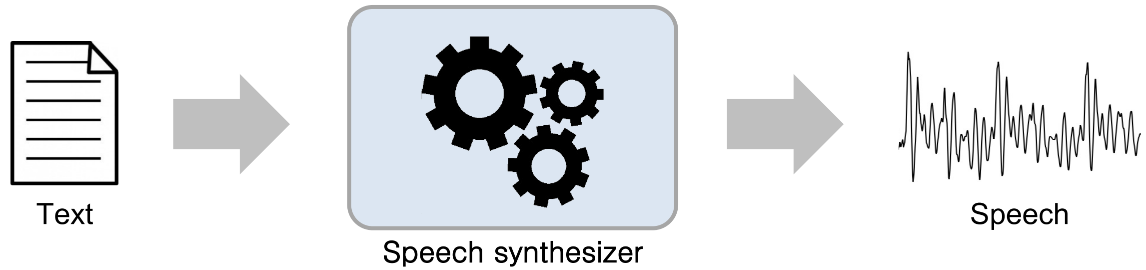
황민제 / HDTS

NAVER CLOVA

CONTENTS

INTRODUCTION

Text-to-Speech (TTS) ?



- The system synthesizing speech waveform from given input text

Application area

Navigation

AI speaker

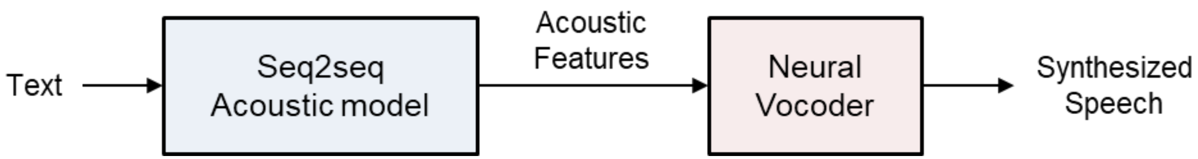
Audiobook

Ai Call

Speech translation

End-to-end TTS system

Concept



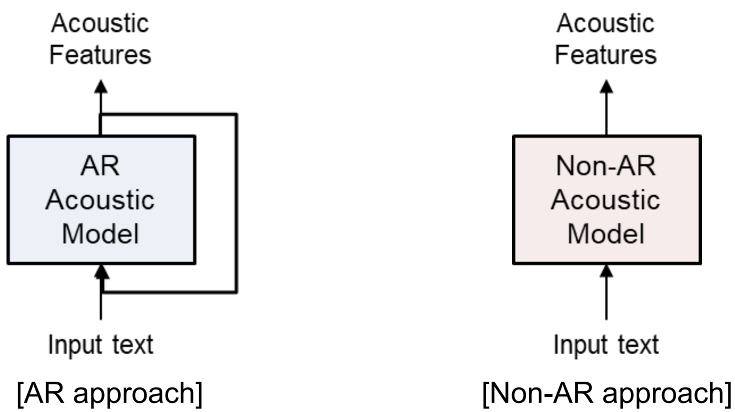
[End-to-end TTS system]

- 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

Acoustic models for end-to-end TTS

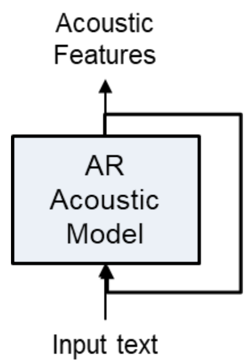
Autoregressive (AR) approach vs. Non-AR approach



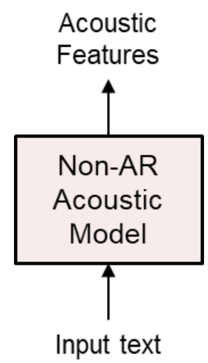
	Autoregressive model	Non-autoregressive model
Generation type	Sequential	Non-sequential
Advantage ☺	High quality	Fast generation
Disadvantage ☹	Slow generation	Not good quality

ACOUSTIC MODELS FOR END-TO-END TTS

Autoregressive (AR) approach vs. Non-AR approach



[AR approach]



[Non-AR approach]

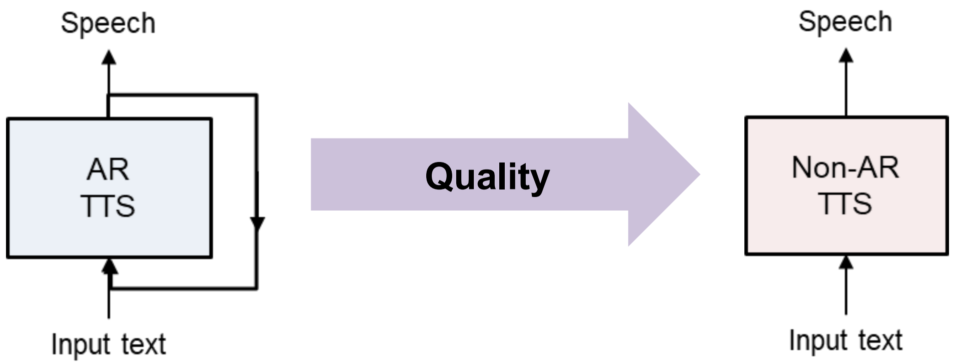
	Autoregressive model	Non-autoregressive model
Generation type	Sequential	Non-sequential
Advantage ☺	High quality	Fast generation
Disadvantage ☹	Slow generation	Not good quality

How could we build the TTS system having both good quality & fast speed?

TTS-DRIVEN DATA AUGMENTATION

Key idea: Let's utilize *data augmentation method*!

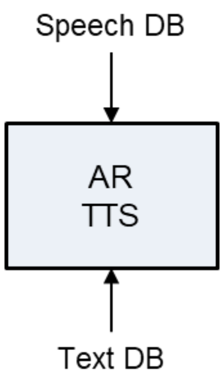
- **Transplant** the quality of AR TTS system to Non-AR TTS system



TTS-DRIVEN DATA AUGMENTATION

Method

Step 1.
Train well designed ***source TTS*** system

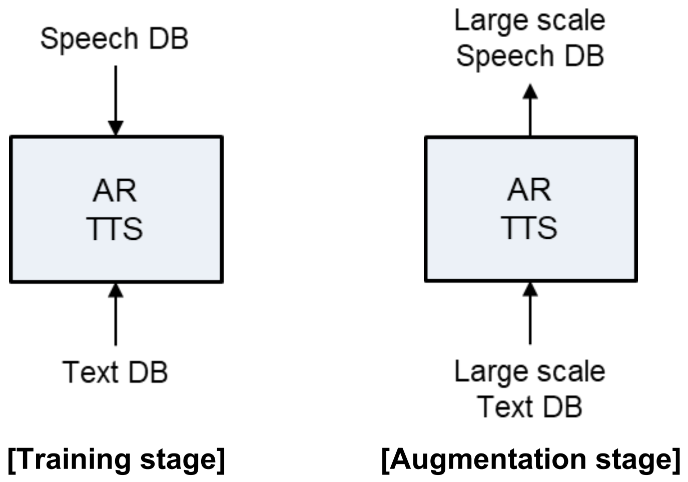


[Training stage]

TTS-DRIVEN DATA AUGMENTATION

Method

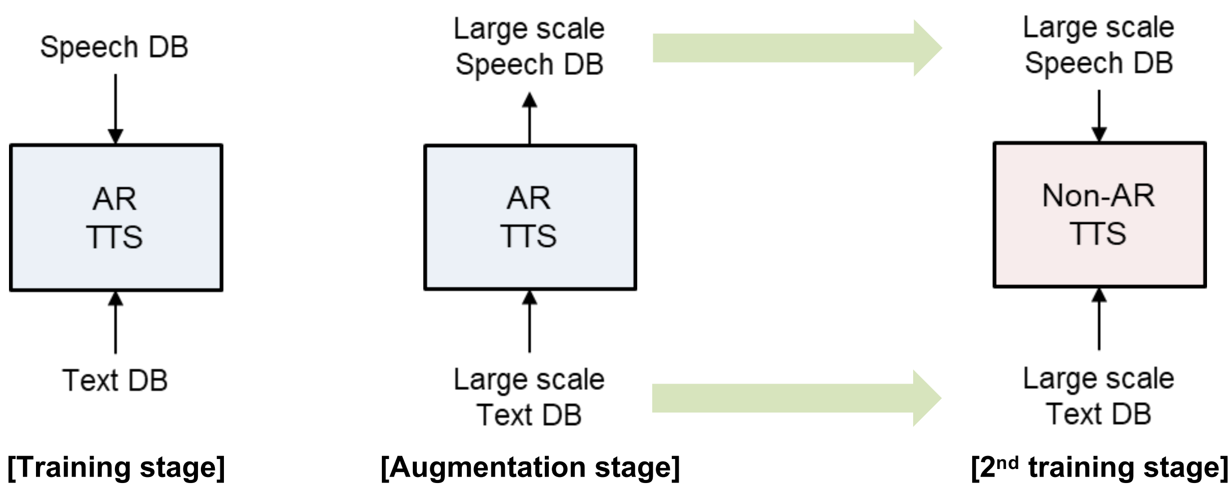
Step 2.
Generate large-scale TTS database



TTS-DRIVEN DATA AUGMENTATION

Method

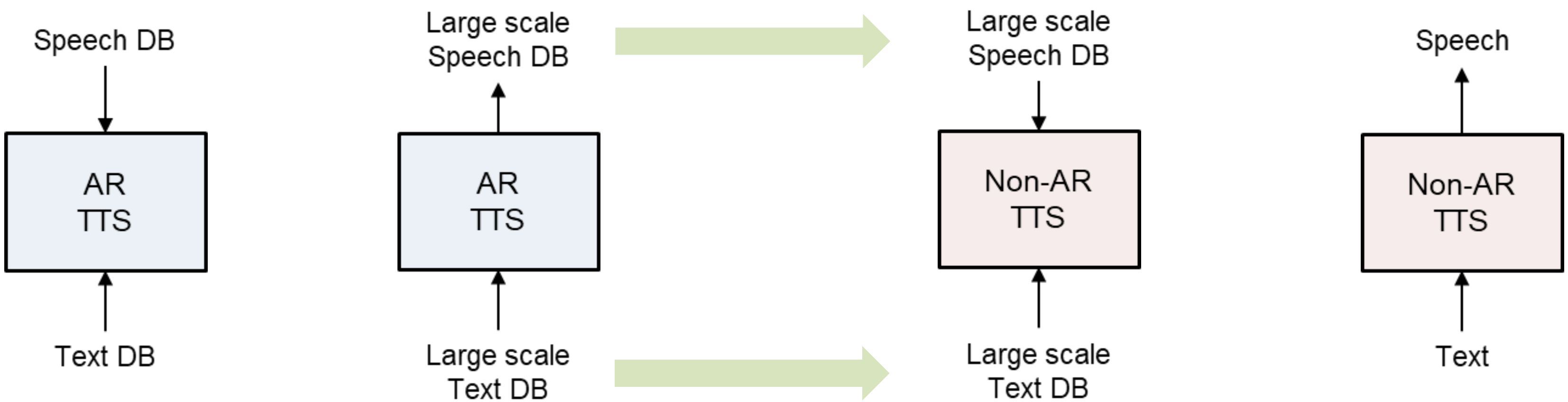
Step 3.
Train ***target TTS*** system by using augmented DB



TTS-DRIVEN DATA AUGMENTATION

Method

Step 4.
Enjoy qualified & fast TTS system ☺



[Training stage] **[Augmentation stage]** **[2nd training stage]** **[Synthesis stage]**

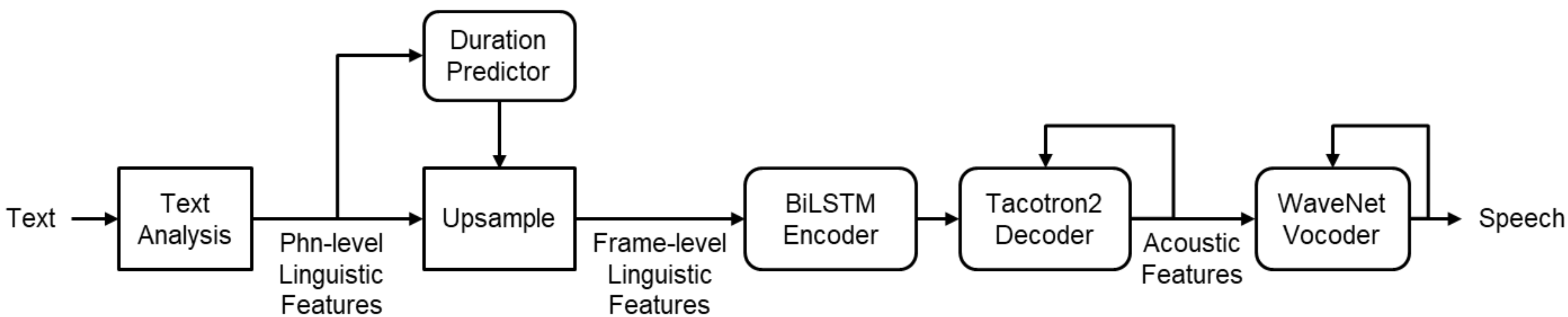
Source TTS system

Requirement

- **Stably** generate **high quality** speech signal

Architecture [1, 4]

- Tacotron2 with duration predictor & linear prediction (LP)-WaveNet vocoder



[1] Shen et. al., “Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions,” in *CoRR*, 2017.
[4] Hwang et. al., " LP-WaveNet: Linear prediction-based wavenet speech synthesis," in *APSIPA*, 2020

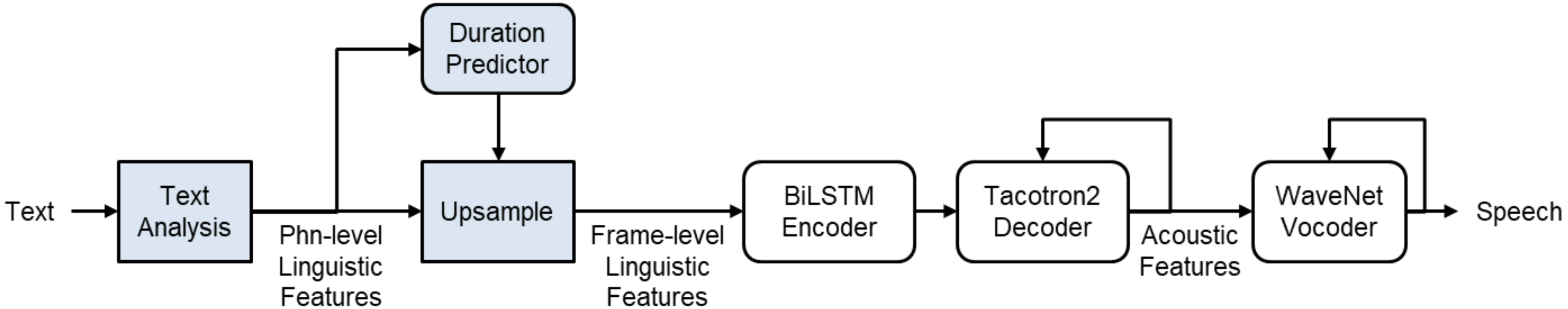
Source TTS system

Requirement

- **Stably** generate **high quality** speech signal

Architecture [1, 4]

- Tacotron2 with duration predictor & linear prediction (LP)-WaveNet vocoder



1. Extract linguistic features (LFs) from input text
2. Estimate duration of each phoneme
3. Upsample LFs to have resolution of acoustic features (AFs)
    - Free from attention failures, e.g., skipping, repetition, collapsing

➡ ***Stable!***

[1] Shen et. al., “Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions,” in *CoRR*, 2017.
[4] Hwang et. al., " LP-WaveNet: Linear prediction-based wavenet speech synthesis," in *APSIPA*, 2020

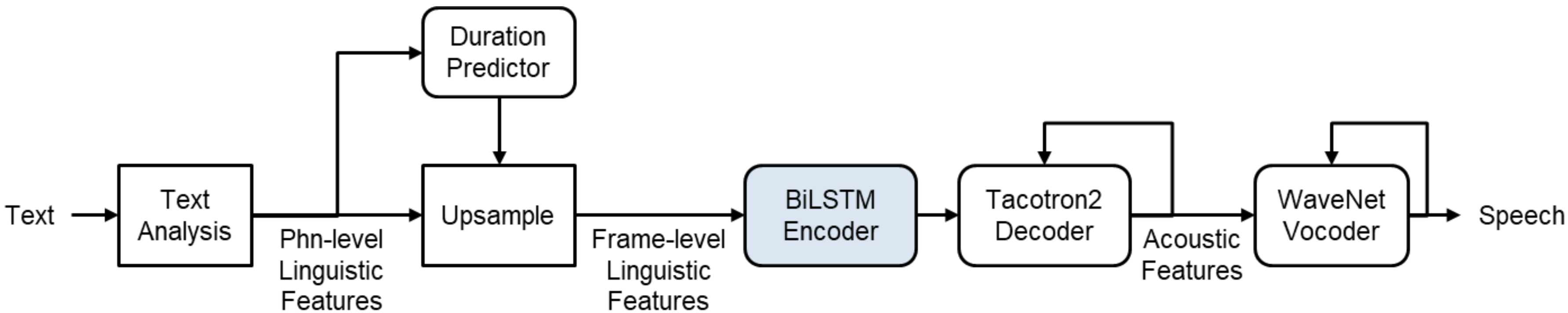
SOURCE TTS SYSTEM

Requirement

- **Stably** generate **high quality** speech signal

Architecture [1, 4]

- Tacotron2 with duration predictor & linear prediction (LP)-WaveNet vocoder



4. Extract high-level context features

[1] Shen et. al., “Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions,” in *CoRR*, 2017.
[4] Hwang et. al., " LP-WaveNet: Linear prediction-based wavenet speech synthesis," in *APSIPA*, 2020

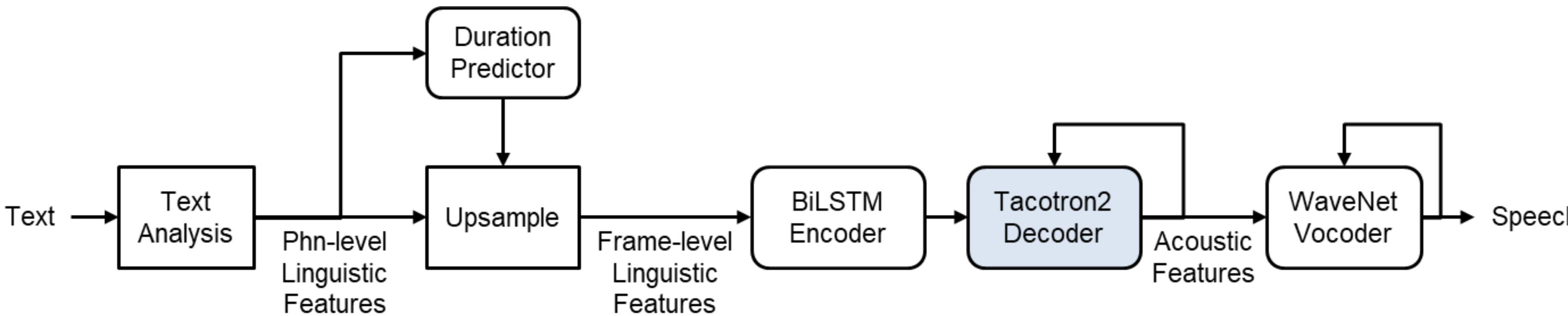
SOURCE TTS SYSTEM

Requirement

- **Stably** generate **high quality** speech signal

Architecture [1, 4]

- Tacotron2 with duration predictor & linear prediction (LP)-WaveNet vocoder



5. Sequentially generate AFs
6. Line spectral frequencies (LSFs) sharpening process
    - Improving spectral clarity of speech signal

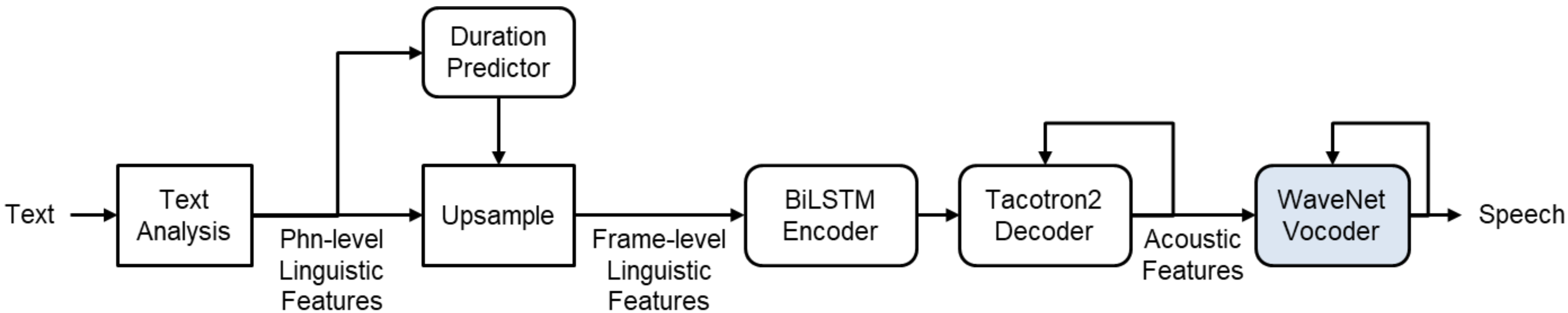
Source TTS system

Requirement

- **Stably** generate **high quality** speech signal

Architecture [1, 4]

- Tacotron2 with duration predictor & linear prediction (LP)-WaveNet vocoder



6. Synthesize speech waveform
7. Distribution sharpening process
   - Reduce randomness of generated waveform

[1] Shen et. al., "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *CoRR*, 2017.
[4] Hwang et. al., " LP-WaveNet: Linear prediction-based wavenet speech synthesis," in *APSIPA*, 2020

TARGET TTS SYSTEM

Requirement

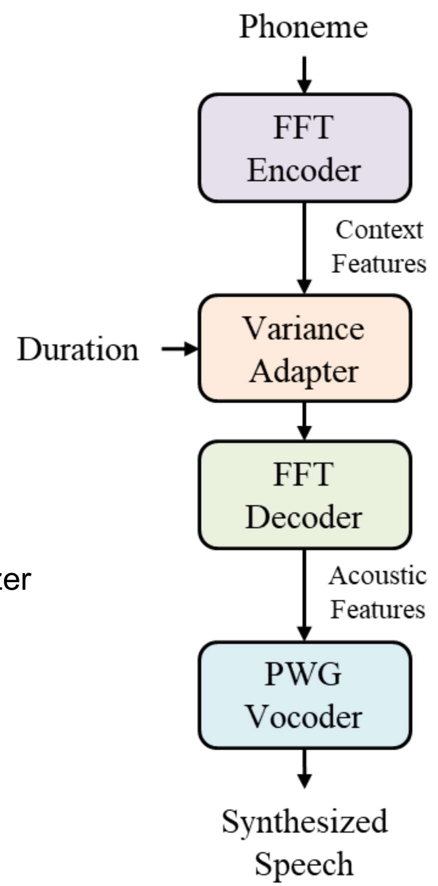
- **Fast synthesis speed**, which is suitable for real-time applications

Acoustic model

- FastSpeech2 [5]
  - Robust to attention errors with fast feature generation speed

Neural vocoder

- Parallel WaveGAN [6]
  - Generative adversarial network (GAN)-based Non-AR waveform synthesizer



[Non-AR TTS for testing]

[5] Yamamoto et. al., “Parallel WaveGan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in *ICASSP*, 2020
[6] Ren et. al., "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Arxiv*, 2020

Summary of TTS systems

Model	Acoustic model	Neural vocoder	Synthesis speed
Source TTS	Tacotron2 w/ duration predictor (AR)	LP-WaveNet (AR)	**578** times **slower** than real time
Target TTS	FastSpeech2 (Non-AR)	Parallel WaveGAN (Non-AR)	**54** times **faster** than real time

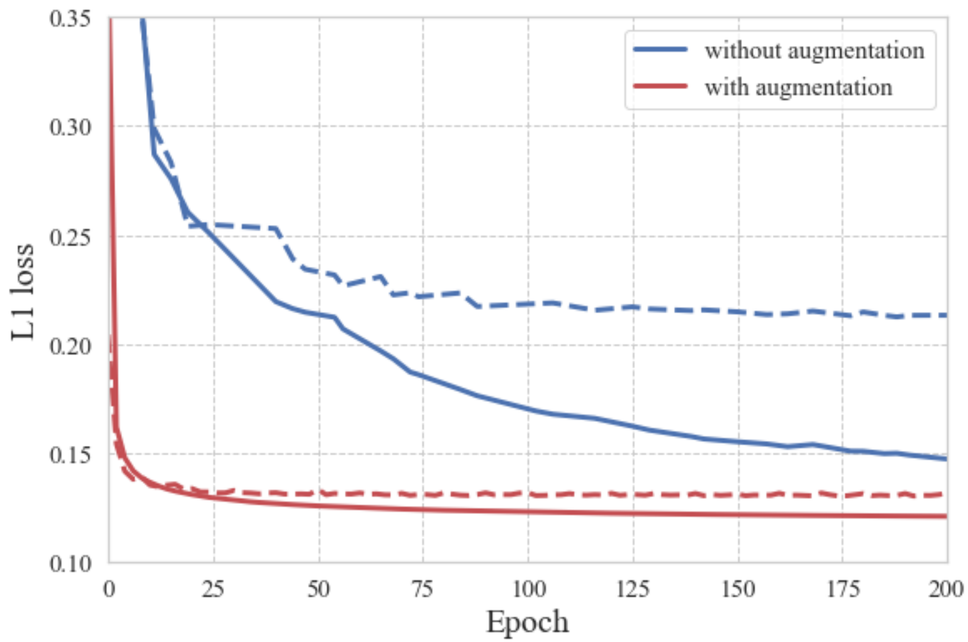
EXPERIMENTS

Database

- Korean female speaker
- Sampling rate / quantization
  - 24-kHz / 16-bit
- Recorded database
  - Train / validation / test : 5 / 1 / 0.5 hours
- Augmented database
  - Train / validation : 170 / 8 hours
- Experimental condition
  - Trained models by using various combination of database
    Case 1) Recorded DB
    Case 2) Augmented DB
    Case 3) Recorded DB + augmented DB

EXPERIMENTS

Loss curve of target FastSpeech2



When the data augmentation is applied...

1. Achieved lower loss value
   ➡ *Better modeling accuracy!*
2. Achieved smaller gap between train / valid set
   ➡ *Better generalization performance!*

EXPERIMENTS

Experiment 1

- Mean opinion score (MOS) test
  - Score the quality of speech (from 1.0 to 5.0)
  - 14 native Korean listeners
  - 20 synthesized utterances from test set

[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

- MOS test results

Model	Analysis / synthesis	Training database	MOS
Recordings	-	-	4.67±0.11
Source TTS	Yes	-	4.19±0.15
	-	Only recorded	4.09±0.17
Target TTS	Yes	-	3.84±0.21
	-	Only recorded	2.70±0.27
	-	Only augmented	3.50±0.27
	-	Both recorded & augmented	3.78±0.23

EXPERIMENTS

[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

Experiment 1

- Mean opinion score (MOS) test
  - Score the quality of speech (from 1.0 to 5.0)
  - 14 native Korean listeners
  - 20 synthesized utterances from test set

- MOS test results

Model	Analysis / synthesis	Training database	MOS
Recordings	-	-	4.67±0.11
Source TTS	Yes	-	4.19±0.15
	-	Only recorded	4.09±0.17
Target TTS	Yes	-	3.84±0.21
	-	Only recorded	2.70±0.27
	-	Only augmented	3.50±0.27
	-	Both recorded & augmented	3.78±0.23

1. ***AR TTS system performed better than Non-AR TTS system.***

EXPERIMENTS

Experiment 1

- Mean opinion score (MOS) test
  - Score the quality of speech (from 1.0 to 5.0)
  - 14 native Korean listeners
  - 20 synthesized utterances from test set

[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

- MOS test results

Model	Analysis / synthesis	Training database	MOS
Recordings	-	-	4.67±0.11
Source TTS	Yes	-	4.19±0.15
	-	Only recorded	4.09±0.17
Target TTS	Yes	-	3.84±0.21
	-	Only recorded	2.70±0.27
	-	Only augmented	3.50±0.27
	-	Both recorded & augmented	3.78±0.23

1. ***AR TTS system performed better than Non-AR TTS system.***
2. ***Perceptual quality of Non-AR TTS was improved when the augmented DB was used.***

EXPERIMENTS

Experiment 1

- Mean opinion score (MOS) test
  - Score the quality of speech (from 1.0 to 5.0)
  - 14 native Korean listeners
  - 20 synthesized utterances from test set

[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

- MOS test results

Model	Analysis / synthesis	Training database	MOS
Recordings	-	-	4.67±0.11
Source TTS	Yes	-	4.19±0.15
	-	Only recorded	4.09±0.17
Target TTS	Yes	-	3.84±0.21
	-	Only recorded	2.70±0.27
	-	Only augmented	3.50±0.27
	-	Both recorded & augmented	3.78±0.23

1. *AR TTS system performed better than Non-AR TTS system.*
2. *Perceptual quality of Non-AR TTS was improved when the augmented DB was used.*
3. ***When both recorded & augmented DB were used, quality of Non-AR TTS reached to analysis / synthesis case.***

EXPERIMENTS

Experiment 2

- Assume enough amount of source DB
  - Use **20 hours** of source DB

- MOS test results

Model	Training database	MOS
Recordings	-	4.67±0.11
Source TTS	Only recorded	4.28±0.16
Target TTS (5hrs)	Only recorded	2.70±0.27
Target TTS (20hrs)	Only recorded	3.32±0.40
	Only augmented	3.58±0.33
	Both recorded & augmented	3.90±0.24

EXPERIMENTS

Experiment 2

- Assume enough amount of source DB
  - Use **20 hours** of source DB

- MOS test results

Model	Training database	MOS
Recordings	-	4.67±0.11
Source TTS	Only recorded	4.28±0.16
Target TTS (5hrs)	Only recorded	2.70±0.27
Target TTS (20hrs)	Only recorded	3.32±0.40
	Only augmented	3.58±0.33
	Both recorded & augmented	3.90±0.24

1. ***When the non-AR TTS is trained by more DB, its perceptual quality was significantly improved.***

EXPERIMENTS

Experiment 2

- Assume enough amount of source DB
  - Use **20 hours** of source DB

- MOS test results

Model	Training database	MOS
Recordings	-	4.67±0.11
Source TTS	Only recorded	4.28±0.16
Target TTS (5hrs)	Only recorded	2.70±0.27
Target TTS (20hrs)	Only recorded	3.32±0.40
	Only augmented	3.58±0.33
	Both recorded & augmented	3.90±0.24

1. ***When the non-AR TTS is trained by more DB, its perceptual quality was significantly improved.***
2. ***However, its quality was still worse than AR-TTS system.***

EXPERIMENTS

Experiment 2

- Assume enough amount of source DB
  - Use **20 hours** of source DB

- MOS test results

Model	Training database	MOS
Recordings	-	4.67±0.11
Source TTS	Only recorded	4.28±0.16
Target TTS (5hrs)	Only recorded	2.70±0.27
Target TTS (20hrs)	Only recorded	3.32±0.40
	Only augmented	3.58±0.33
	Both recorded & augmented	3.90±0.24

1. *When the non-AR TTS is trained by more DB, its perceptual quality was significantly improved.*
2. *However, its quality was still worse than AR-TTS system.*
3. ***Similar with the case of 5 hours, Non-AR TTS could be improved by applying data augmentation method.***

SPEECH SAMPLES

Recorded

Source TTS (only natural DB)

Target TTS (only natural DB)

Target TTS (only augmented DB)

Target TTS (both recorded DB & augmented DB)

Summary & Conclusion

Summary

- Proposed a TTS-driven data augmentation method for fast & high quality TTS system

Autoregressive (AR) TTS vs. Non-AR TTS

- AR TTS : High quality, but slow synthesis speed
- Non-AR TTS : Fast synthesis speed, but low quality

TTS-driven data augmentation method

- Augment TTS DB by using high quality AR TTS system
- Train Non-AR TTS system by using augmented DB for improving its quality

Performance evaluation results

- Significantly improved the performance of Non-AR TTS system
  - From 2.70 MOS to 3.78 MOS by using only 5 hours training DB

Thank you!