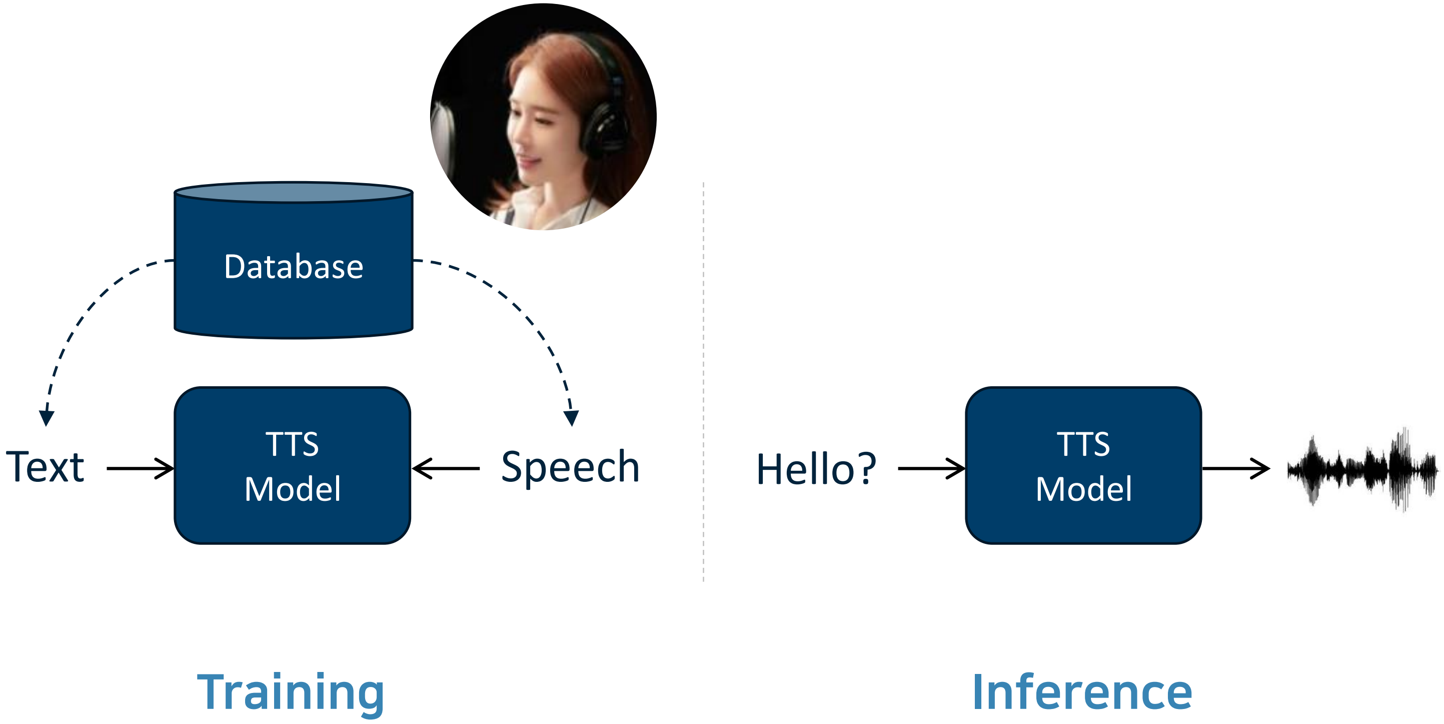
Speech synthesis and its applications

Eunwoo Song / Naver Cloud

Introduction

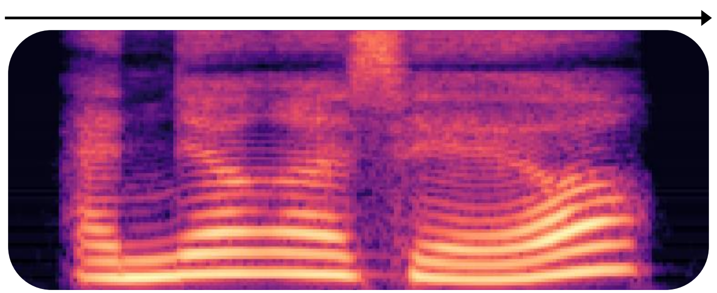
Deep learning-based TTS system



Human-like voice quality ☺

Introduction

Deep learning-based TTS system



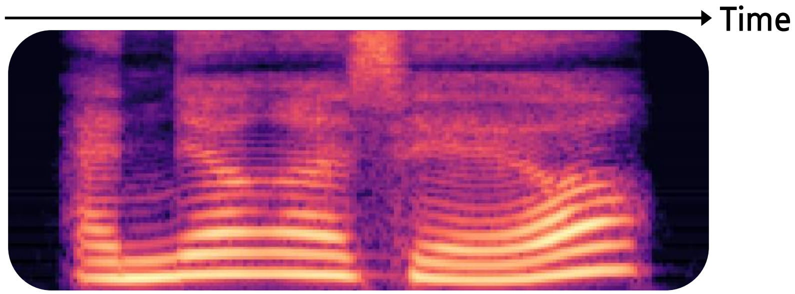
Text → Acoustic Model → Acoustic Parameters → Vocoding Model →

Estimating acoustic parameters from text inputs

DNN TTS = Acoustic model + Vocoding model

Introduction

Deep learning-based TTS system



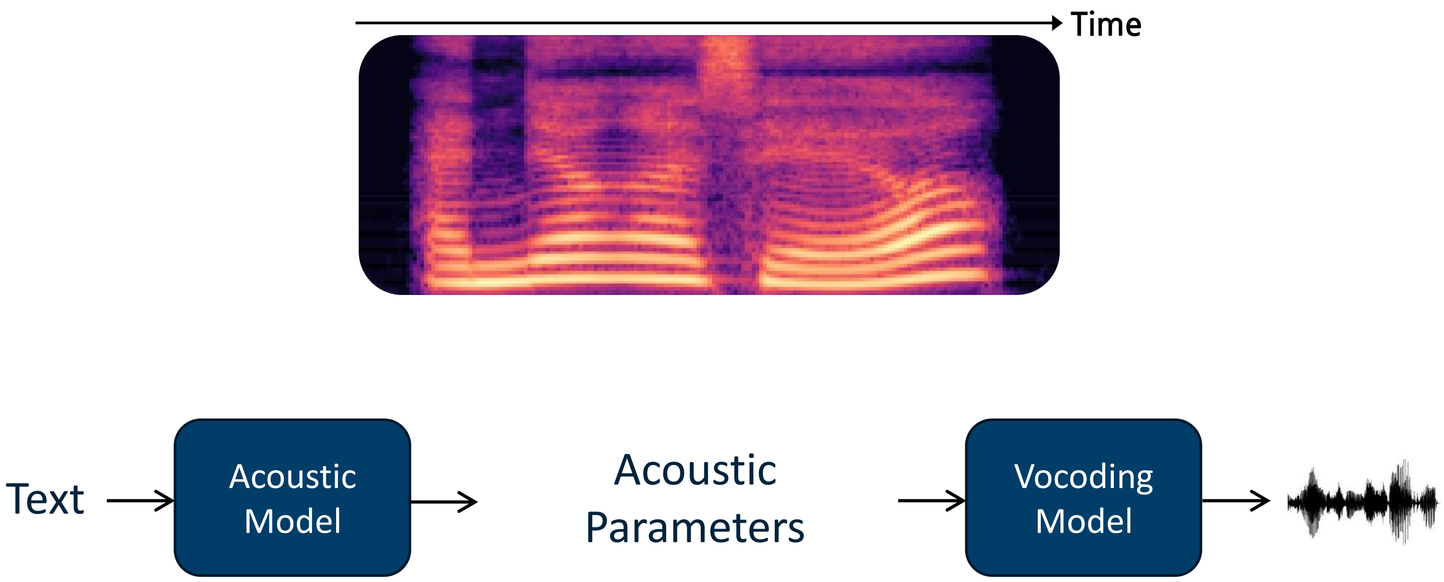
Text → Acoustic Model → Acoustic Parameters → Vocoding Model → 

Estimating speech signals from acoustic parameters

DNN TTS = Acoustic model + Vocoding model

Introduction

Deep learning-based TTS system



본 강의에서는 Acoustic Model 과 Vocoder 기술 정리를 통해

딥러닝 기반의 음성 합성 시스템에 대한 이해도를 높이고자 합니다.

Speech synthesis and its applications

1. Speech analysis: Mel-spectrogram
2. Acoustic models: From text to acoustic parameters
3. Vocoder: From acoustic parameters to speech
4. Fully end-to-end speech synthesis
5. Applications

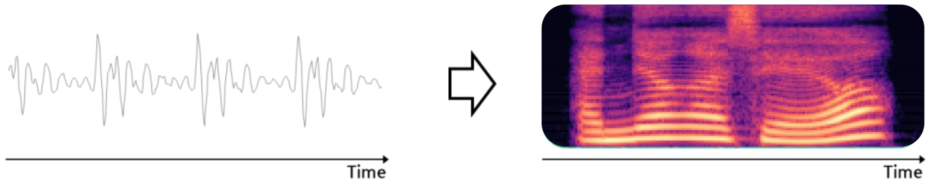
Speech analysis

Overview

Text → Acoustic Model → Acoustic Parameters → Vocoding Model → 

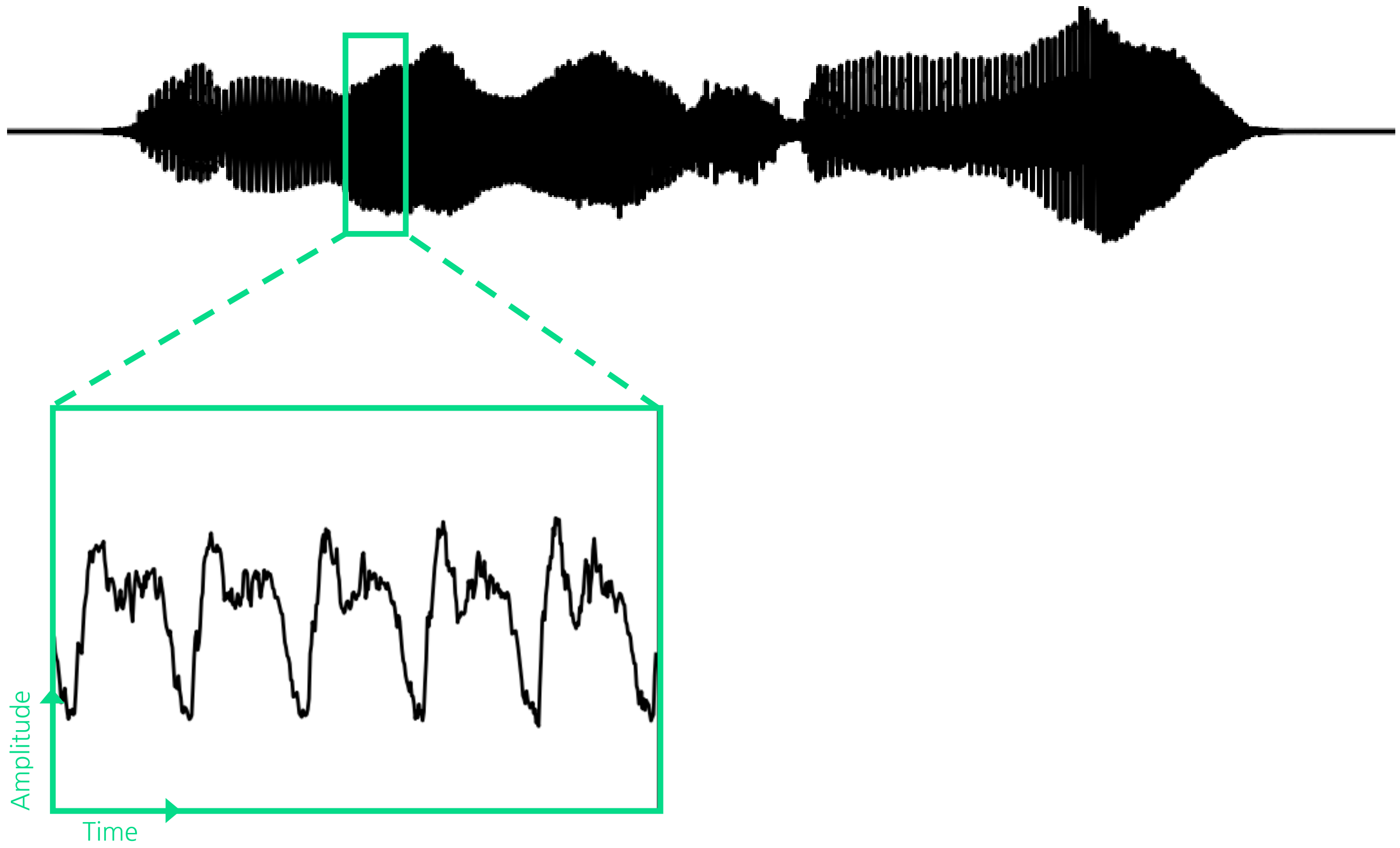
Acoustic parameters..?

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



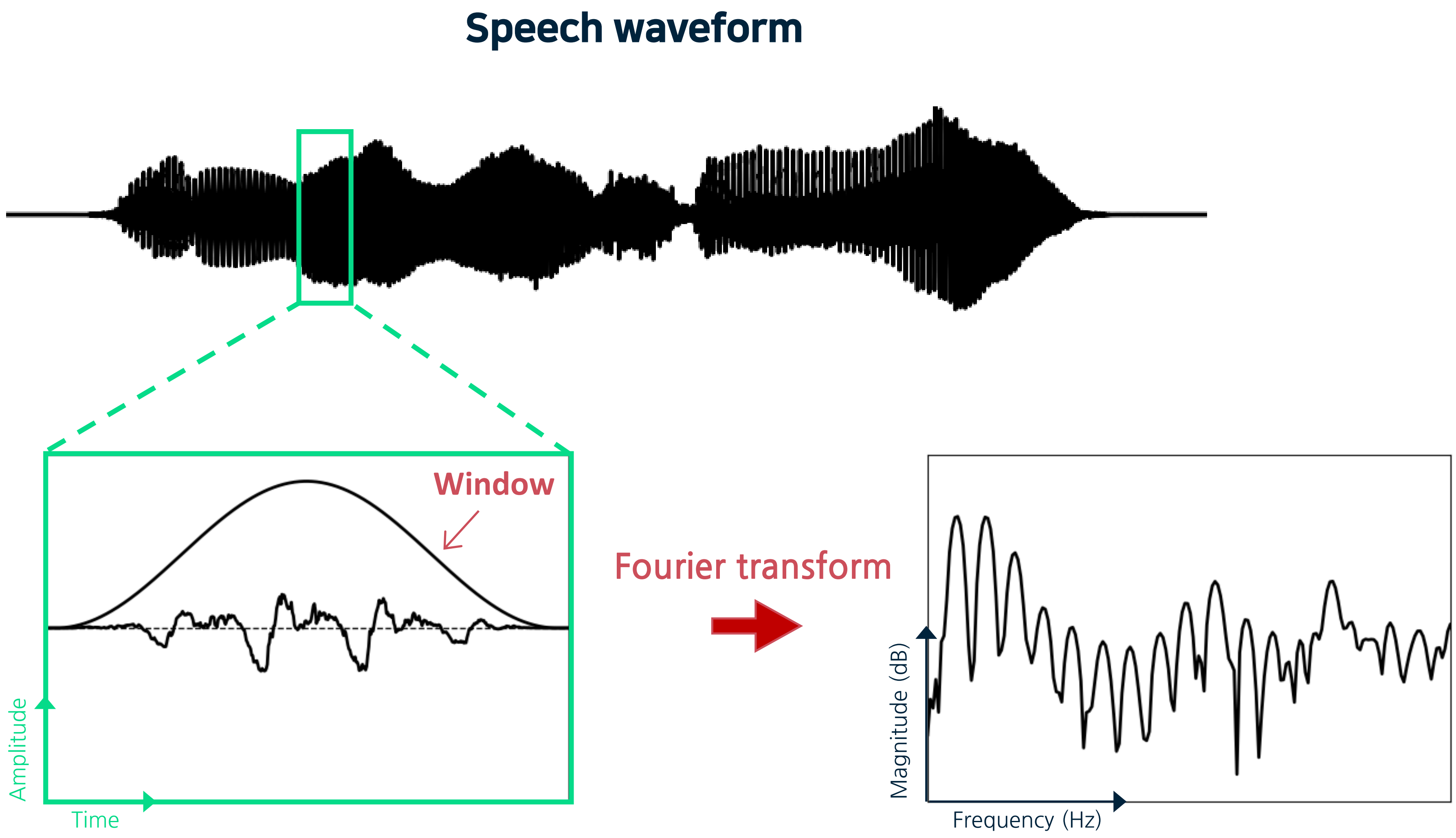
Speech analysis

Speech waveform



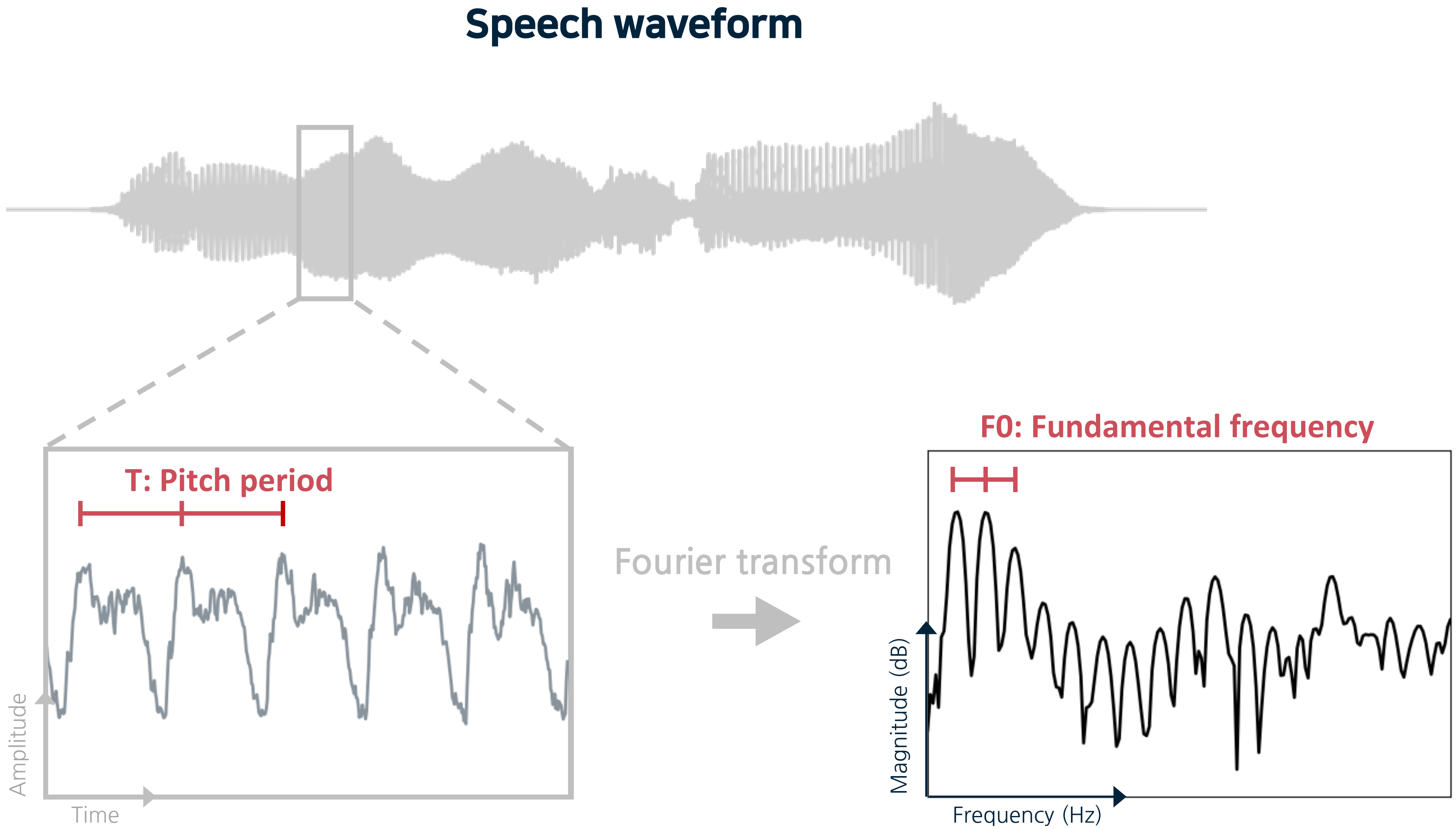
음성 신호는 시간 축에서 특정한 에너지를 갖는 파형의 형태로 존재합니다

Speech analysis



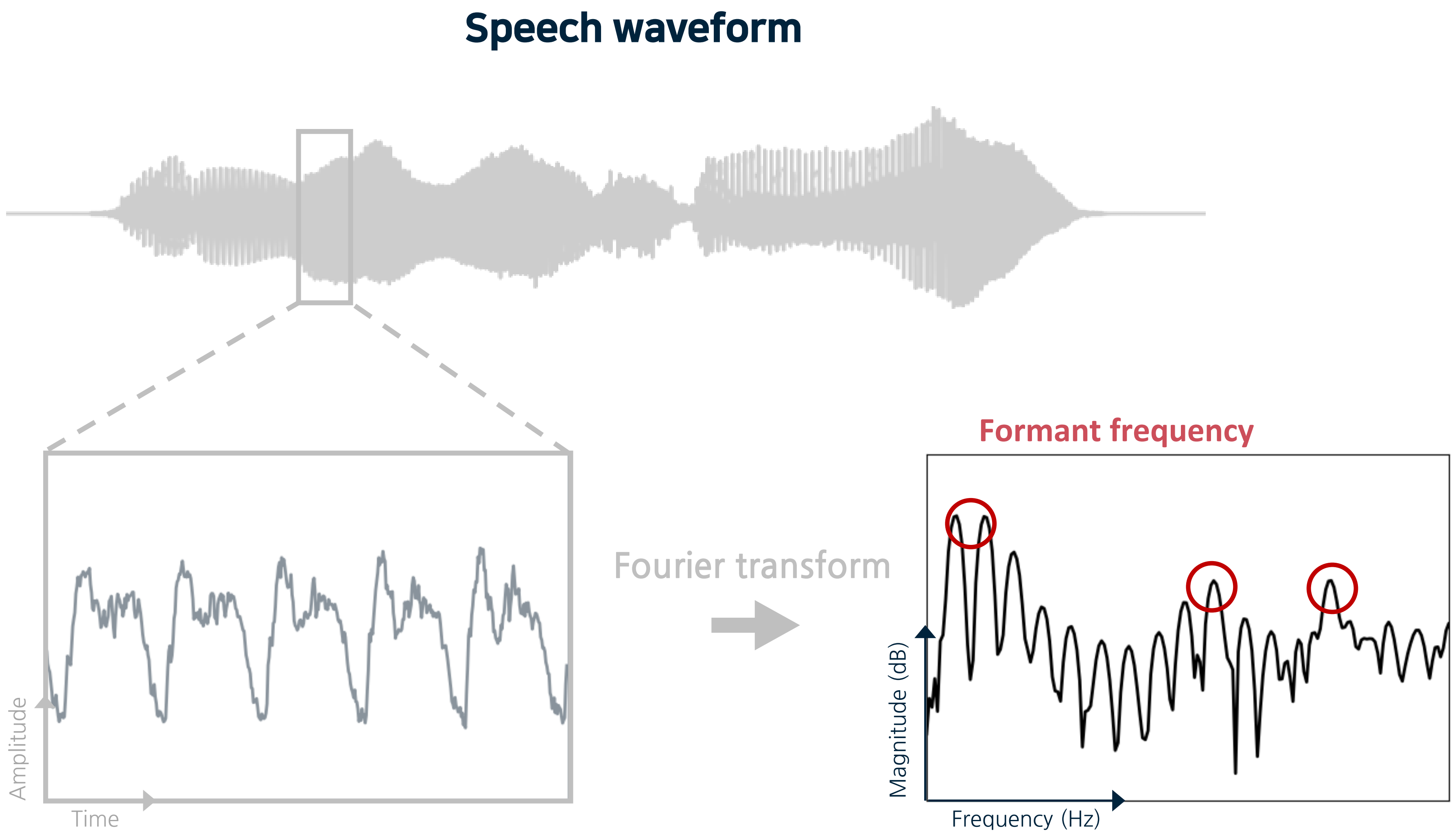
Fourier 변환을 통해 주파수 축에서 음성을 관찰할 수 있습니다

Speech analysis



F0 의 높낮이에 따라 목소리의 톤이 결정됩니다 (아↘아↗)

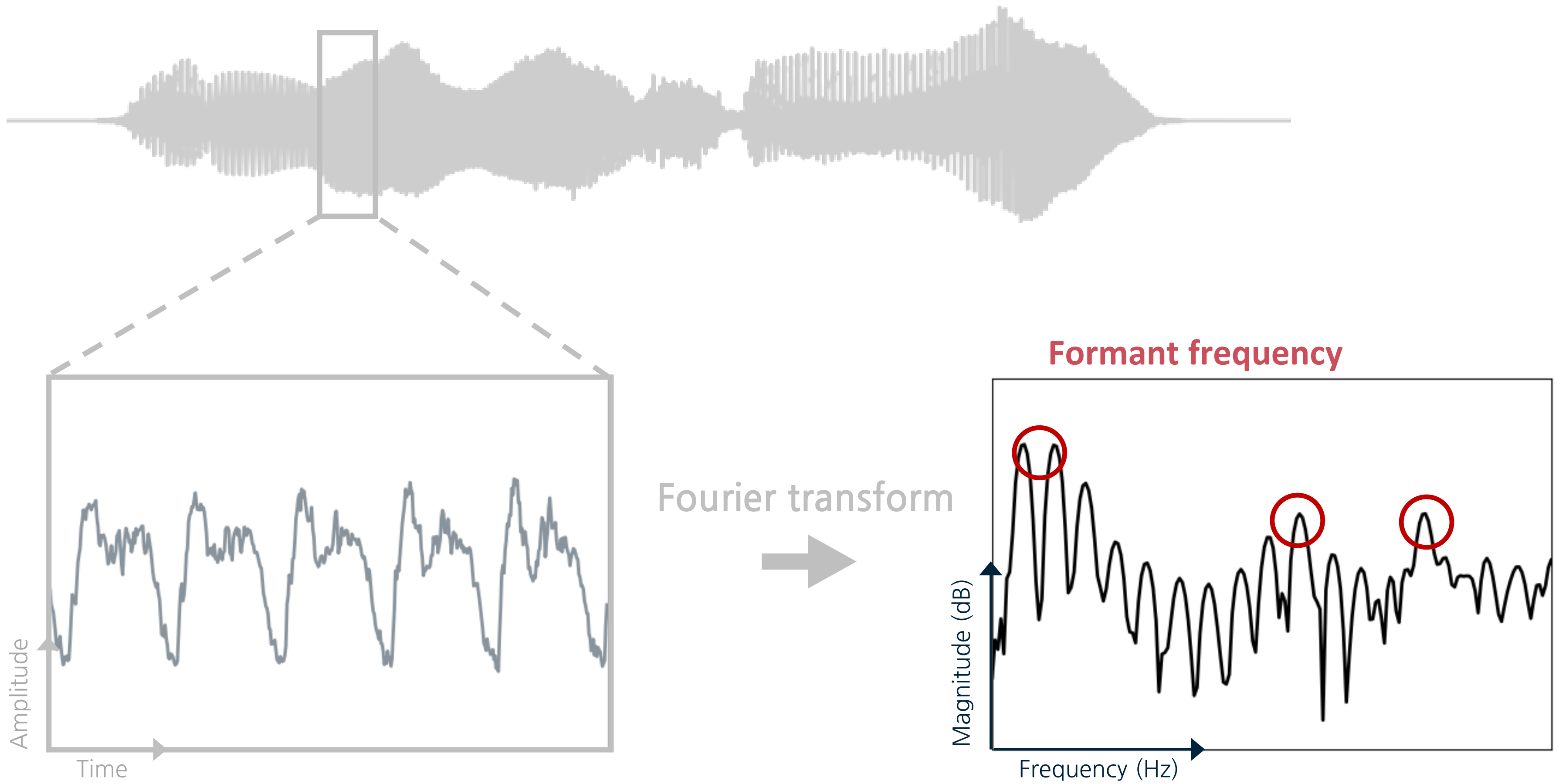
Speech analysis



높은 에너지를 갖는 (spectral peak) 주파수 성분을 formant frequency 라고 정의합니다

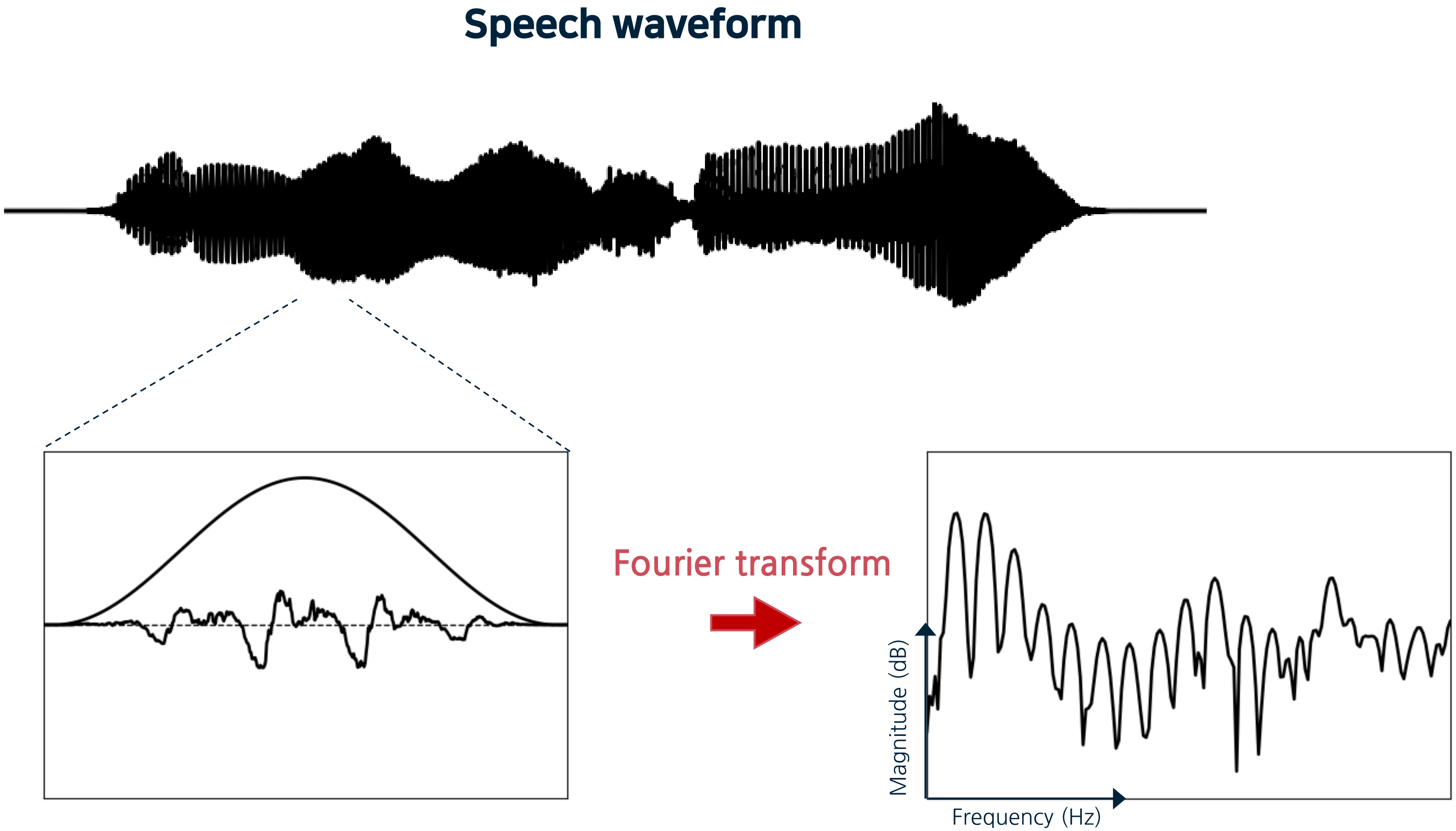
Speech analysis

Speech waveform

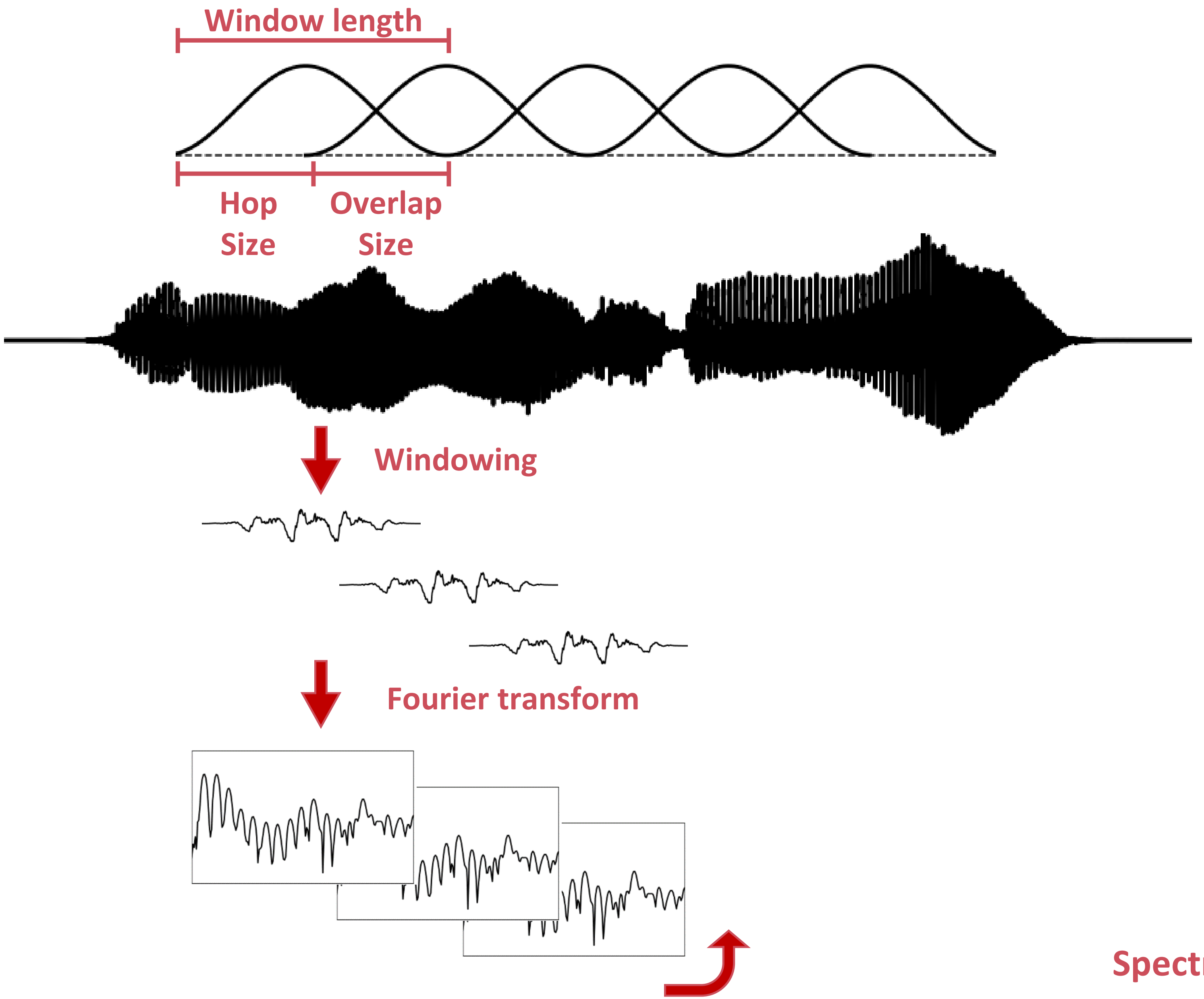


Formant 의 위치에 따라 발음이 결정됩니다 (아/에/이/오/우)

Speech analysis



복잡해 보이는 시간 축 신호를 주파수 축에서 보면 음성을 분석하기 쉬워집니다

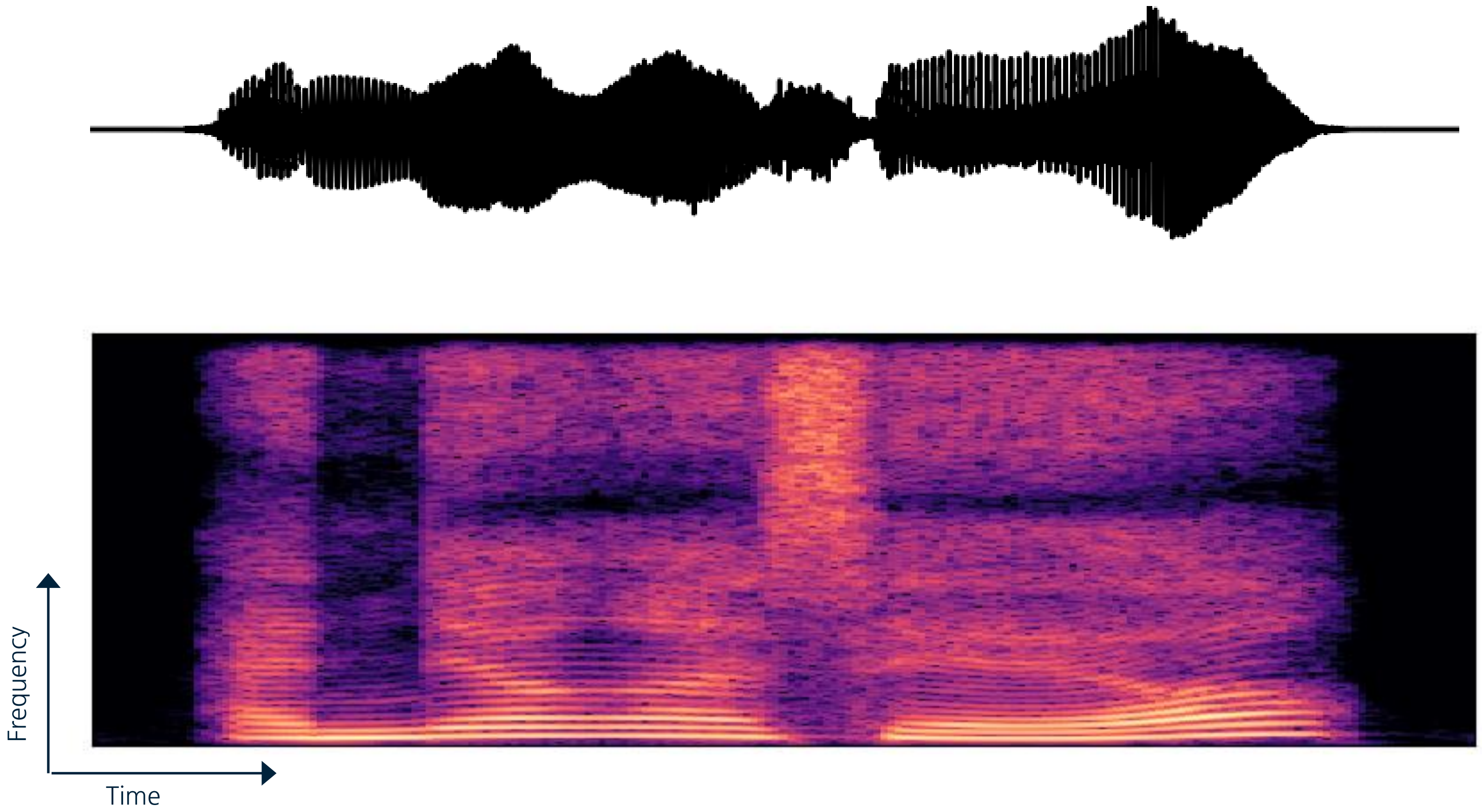


Spectrogram

STFT 신호를 시간 축으로 붙인 2D 이미지

Speech analysis

Spectrogram

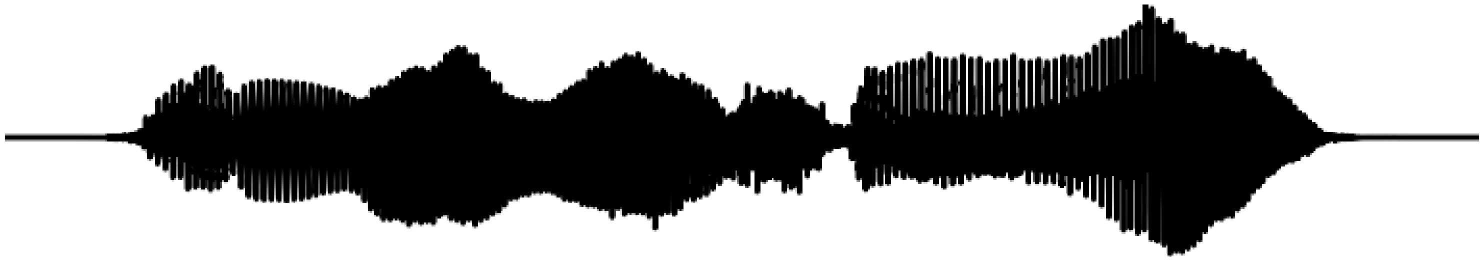


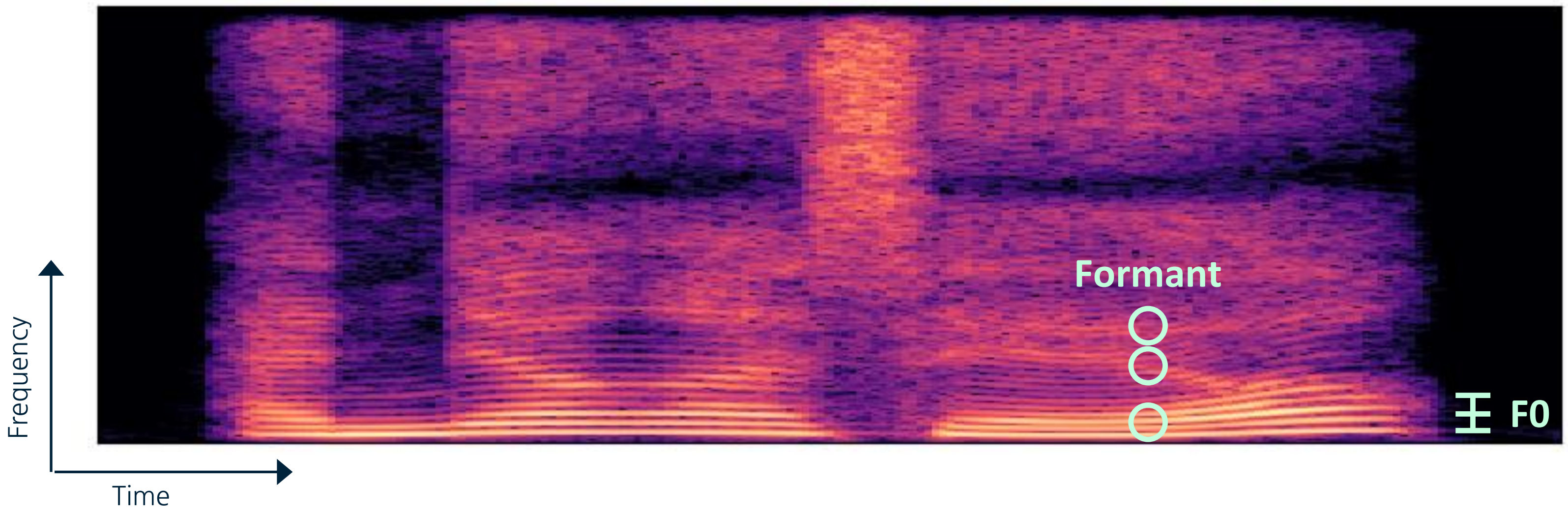
Spectrogram

STFT 신호를 시간 축으로 붙인 2D 이미지

Speech analysis

Spectrogram



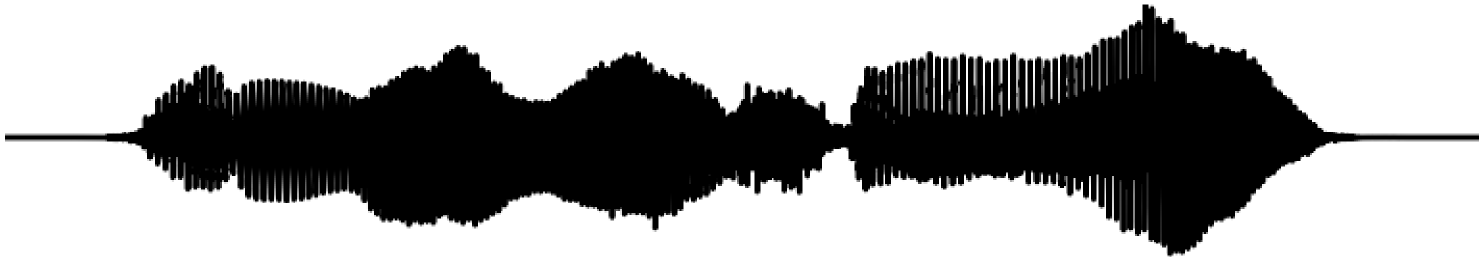


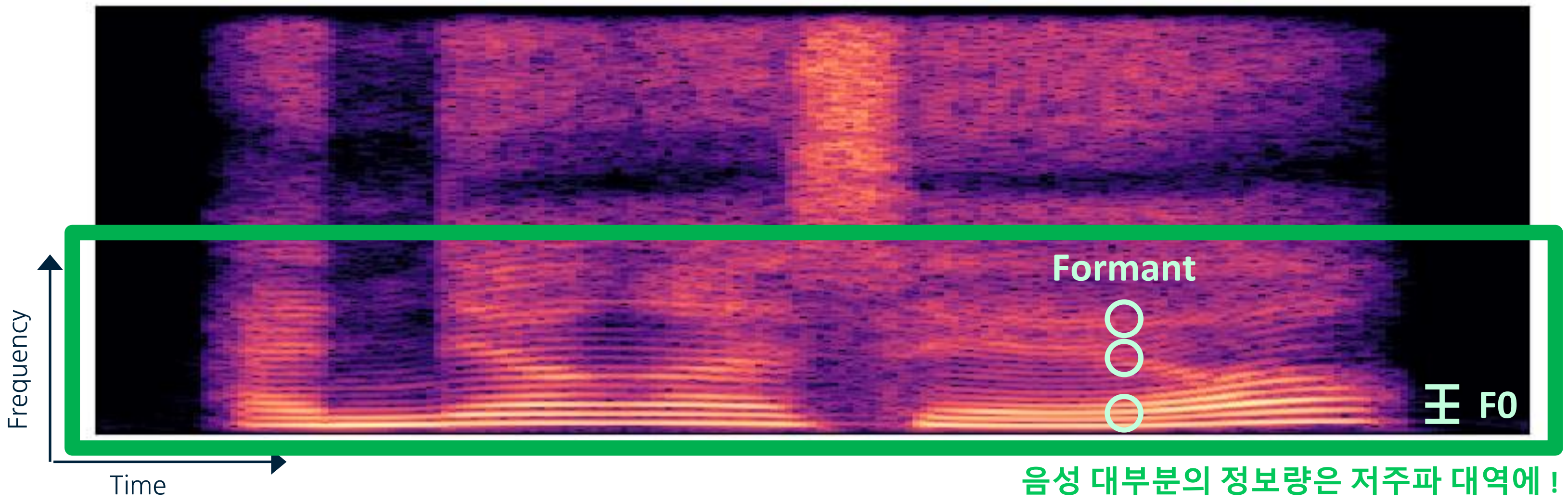
Spectrogram

음성을 시간-주파수 축에서 분석할 수 있게 되었습니다

Speech analysis

Spectrogram





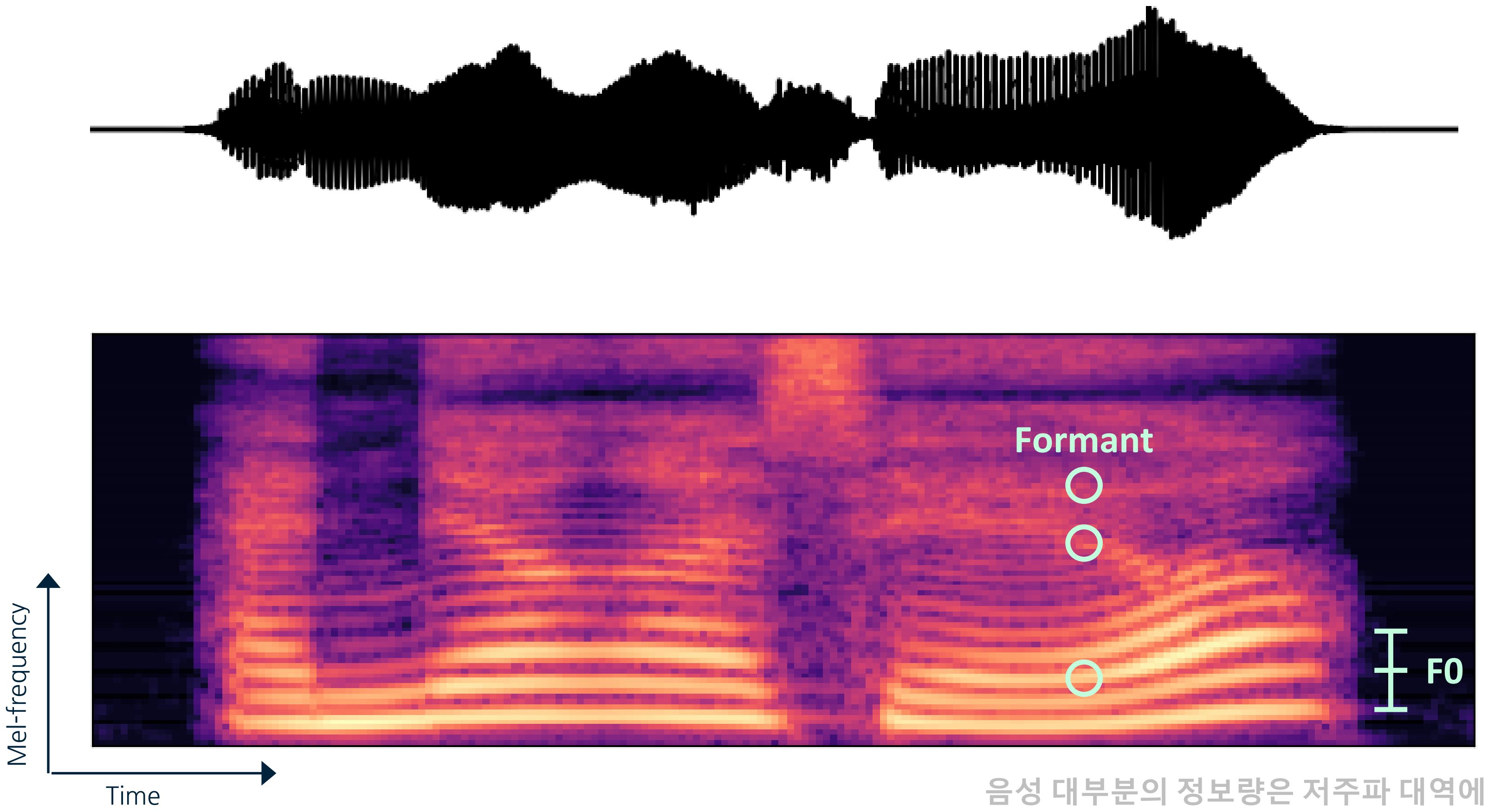
음성 대부분의 정보량은 저주파 대역에 !

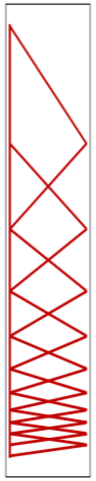
저주파 대역의 정보량에 집중할 수 있다면?

음성을 시간-주파수 축에서 분석할 수 있게 되었습니다

Speech analysis

Mel-spectrogram





주파수 축으로

Mel-filterbank 적용

음성 대부분의 정보량은 저주파 대역에 !

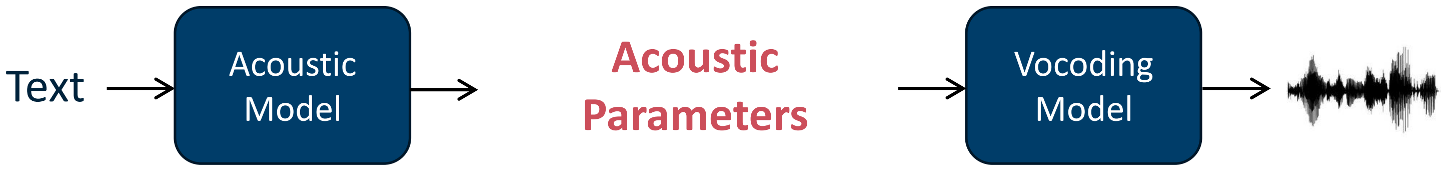
저주파 대역의 정보량에 집중할 수 있다면?

모델이 음성 신호를 이해하기 쉬워집니다 ← 음성을 시간-주파수 축에서 분석을 더 잘 할 수 있습니다

Speech analysis

Mel-spectrogram

Acoustic model 과 **vocoder** 를 연결하는 **매개체** 역할을 하는 것이 **Mel-spectrogram**



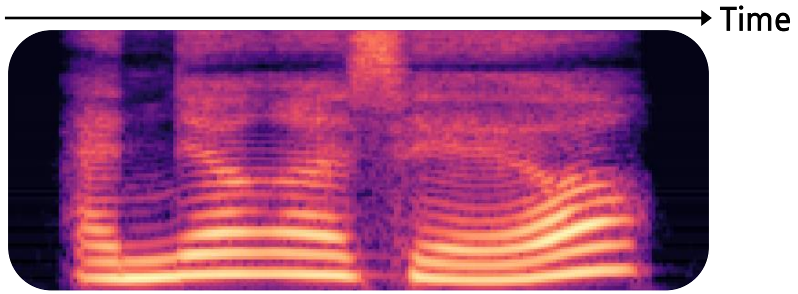
DNN TTS = Acoustic model + Vocoding model

Speech synthesis and its applications

1. Speech analysis: Mel-spectrogram
2. Acoustic models: From text to acoustic parameters
3. Vocoder: From acoustic parameters to speech
4. Fully end-to-end speech synthesis
5. Applications

Acoustic model

Estimating **acoustic parameters** from **text** inputs



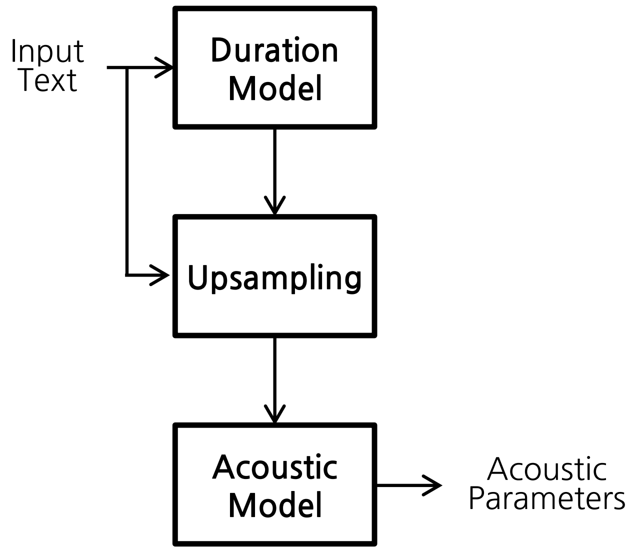
Estimating acoustic parameters from text inputs

Acoustic model

Estimating **acoustic parameters** from **text** inputs

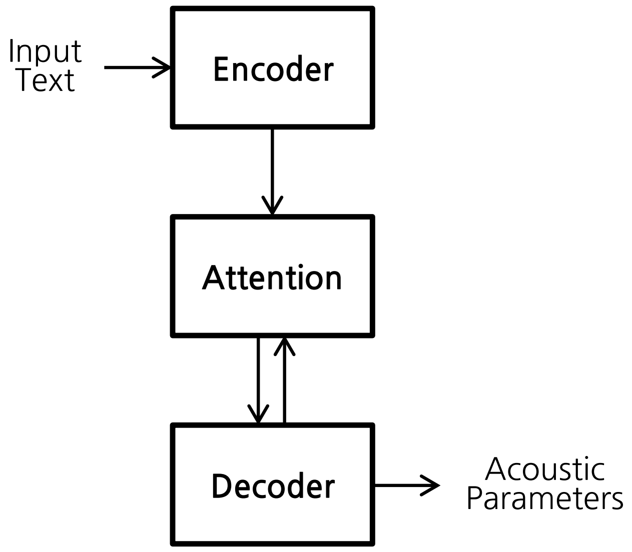
Statistical parametric speech synthesis

- Simple deep learning model (FF+LSTM)



End-to-end speech synthesis

- Seq2seq model



Acoustic model

Statistical parametric speech synthesis

STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING DEEP NEURAL NETWORKS

Heiga Zen, Andrew Senior, Mike Schuster

Google

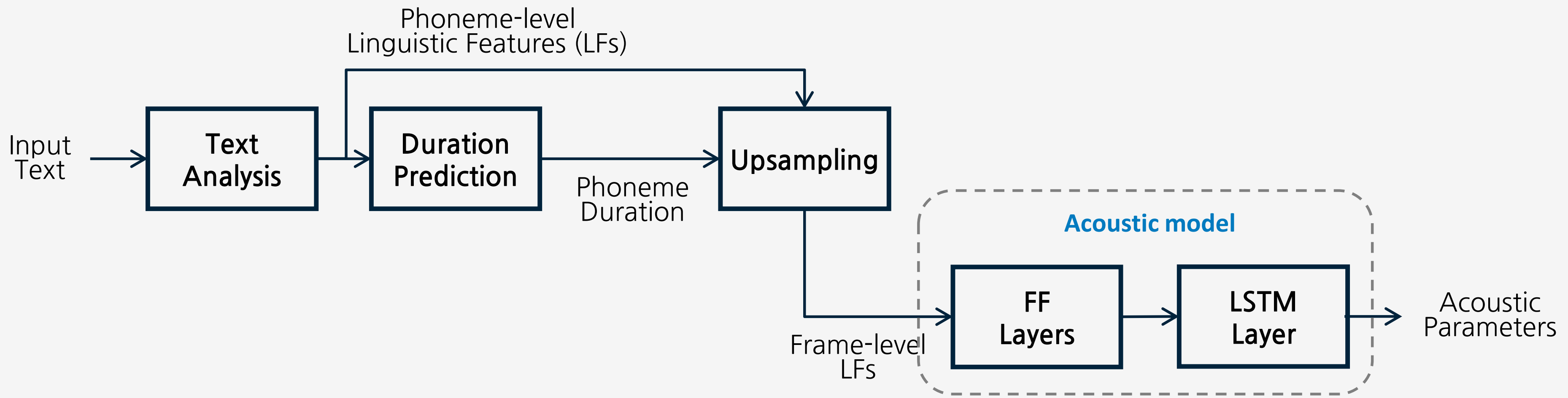
{heigazen,andrewsenior,schuster}@google.com

ABSTRACT

Conventional approaches to statistical parametric speech synthesis typically use decision tree-clustered context-dependent hidden Markov models (HMMs) to represent probability densities of speech parameters given texts. Speech parameters are generated from the probability densities to maximize their output probabilities, then a speech waveform is reconstructed from the generated parameters. This approach is reasonably effective but has a couple of limitations, *e.g.* decision trees are inefficient to model complex context dependencies. This paper examines an alternative scheme that is based on a deep neural network (DNN). The relationship between input texts and their acoustic realizations is modeled by a DNN. The use of the DNN can address some limitations of the conventional approach. Experimental results show that the DNN-based systems outperformed the HMM-based systems with similar numbers of parameters.

Acoustic model

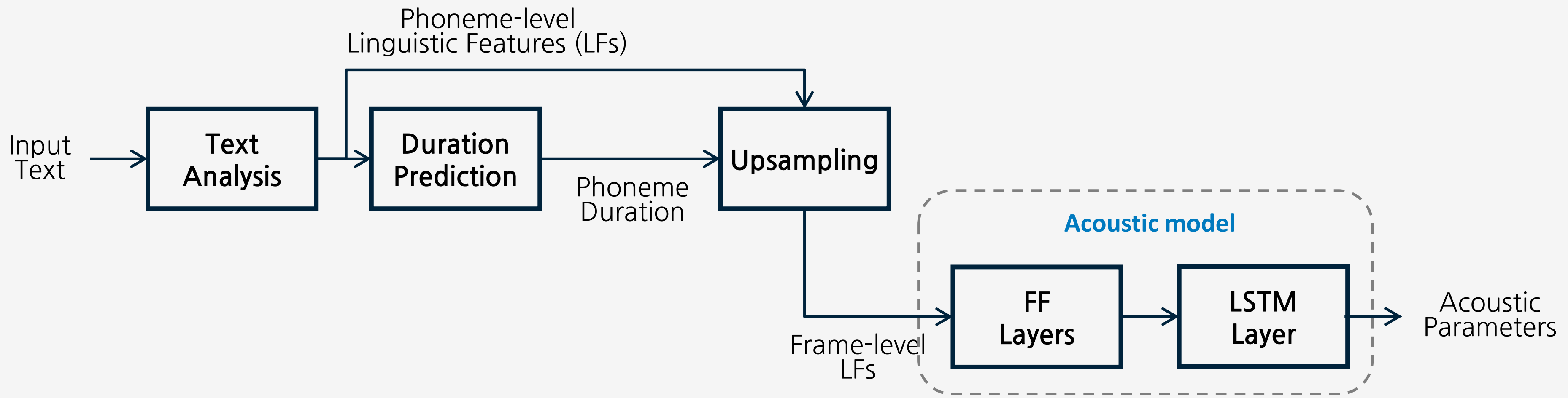
Statistical parametric speech synthesis



Simple and compact	가볍다 + 빠르다
1:1 mapping between linguistic and acoustic features	안정적이다

Acoustic model

Statistical parametric speech synthesis



Simple and compact

1:1 mapping between linguistic and acoustic features

합성음 **품질**이 좋지 않다

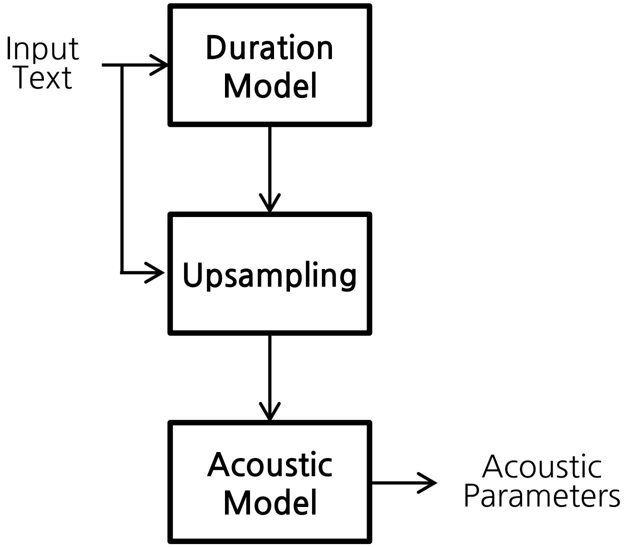
Phoneme segmentation 을 위한 **비용**이 많이 든다

Acoustic model

Estimating **acoustic parameters** from **text** inputs

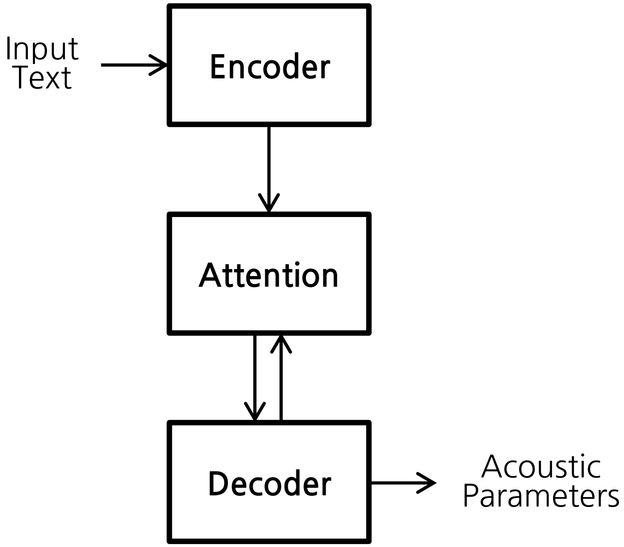
Statistical parametric speech synthesis

- Simple deep learning model (FF+LSTM)



End-to-end speech synthesis

- Seq2seq model



Acoustic model

Tacotron 2

NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS

Jonathan Shen[1], *Ruoming Pang*[1], *Ron J. Weiss*[1], *Mike Schuster*[1], *Navdeep Jaitly*[1], *Zongheng Yang*[*2], *Zhifeng Chen*[1], *Yu Zhang*[1], *Yuxuan Wang*[1], *RJ Skerry-Ryan*[1], *Rif A. Saurous*[1], *Yannis Agiomyrgiannakis*[1], *and Yonghui Wu*[1]

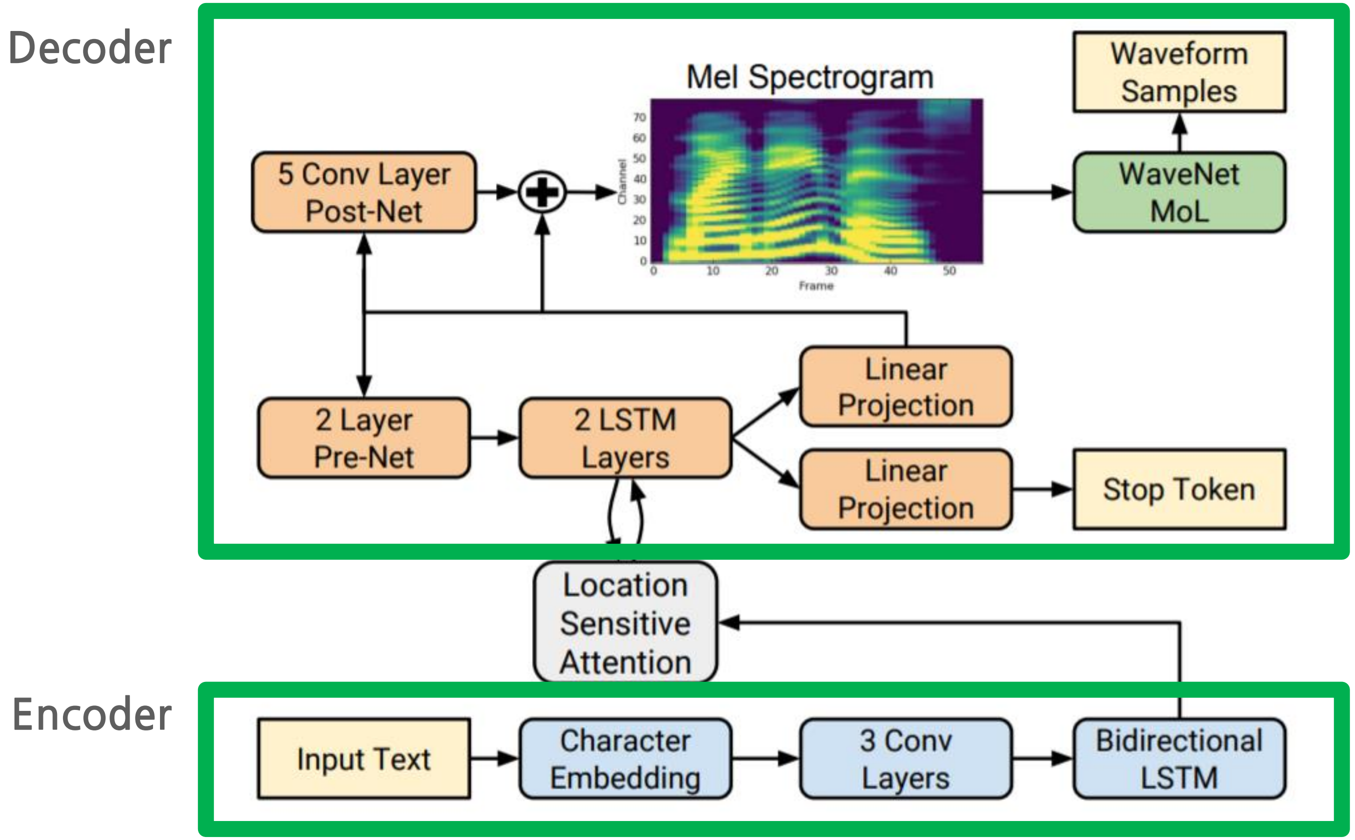
[1]Google, Inc., [2]University of California, Berkeley,
{jonathanasdf,rpang,yonghui}@google.com

ABSTRACT

This paper describes Tacotron 2, a neural network architecture for speech synthesis directly from text. The system is composed of a recurrent sequence-to-sequence feature prediction network that maps character embeddings to mel-scale spectrograms, followed by a modified WaveNet model acting as a vocoder to synthesize time-domain waveforms from those spectrograms. Our model achieves a mean opinion score (MOS) of 4.53 comparable to a MOS of 4.58 for professionally recorded speech. To validate our design choices, we present ablation studies of key components of our system and evaluate the impact of using mel spectrograms as the conditioning input to WaveNet instead of linguistic, duration, and F_0 features. We further show that using this compact acoustic intermediate representation allows for a significant reduction in the size of the WaveNet architecture.

Acoustic model

Tacotron 2



Acoustic model

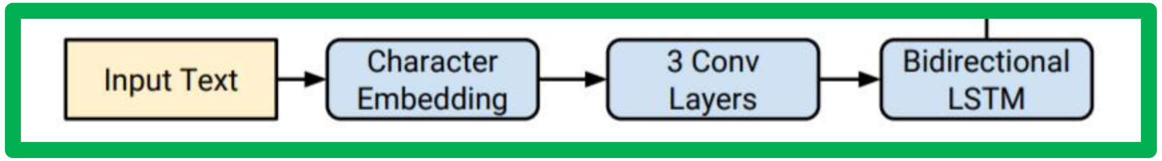
Tacotron 2

Input: Linguistic feature 가 아닌 character embedding

또는 phoneme

대신 Conv. + LSTM 모듈을 이용해 high-level context feature 를 얻어낼 수 있음

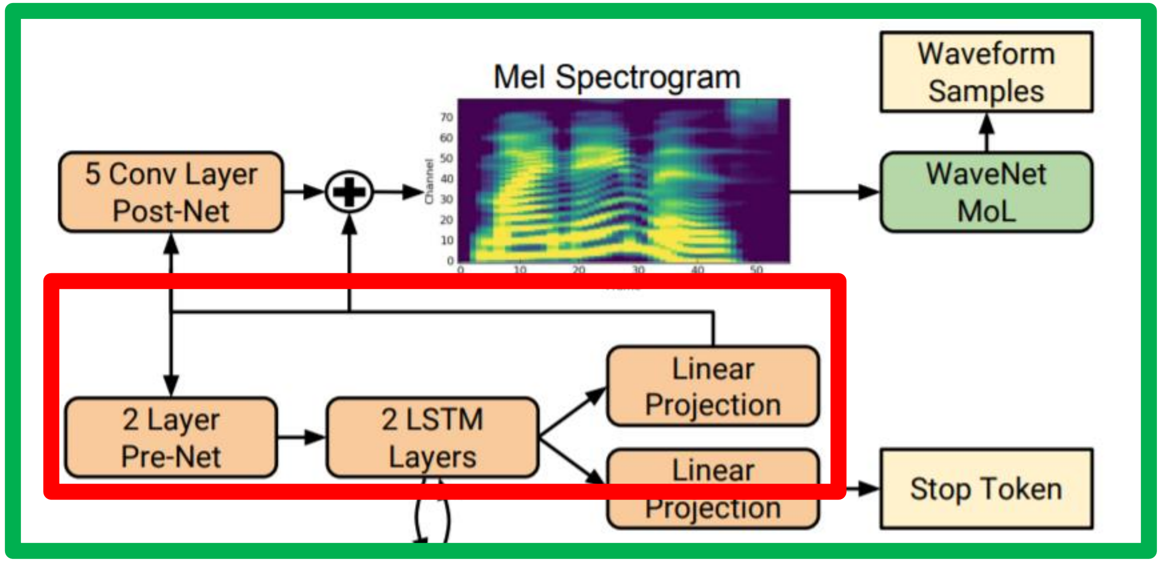
Encoder



Acoustic model

Tacotron 2

Decoder

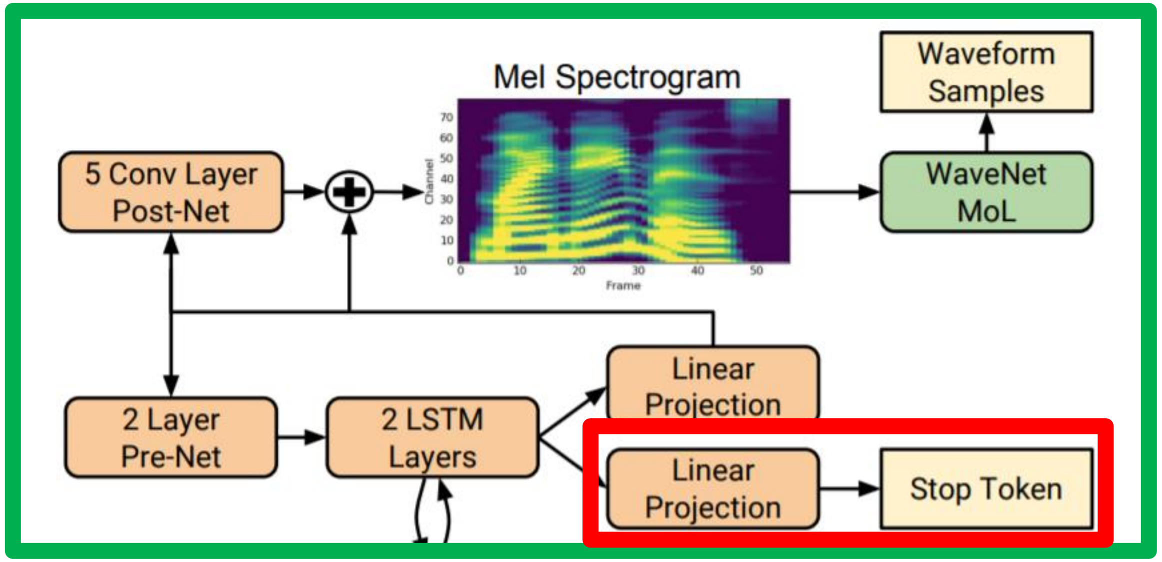


Autoregressive decoder: 합성음 품질을 높임

Acoustic model

Tacotron 2

Decoder



Autoregressive decoder: 합성음 품질을 높임

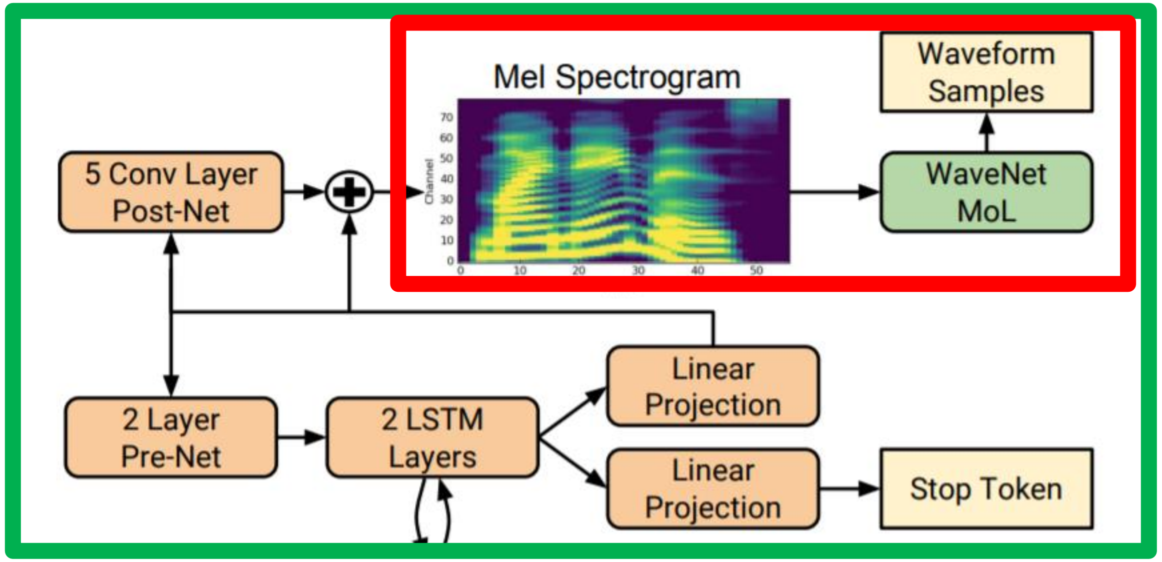
Stop token: 발화의 종료 시점을 추정할 수 있음

cf. Tacotron 1: 발화 종료와 상관 없이 일정 길이 만큼 음성을 생성해야 했음

Acoustic model

Tacotron 2

Decoder



Autoregressive decoder: 합성음 품질을 높임

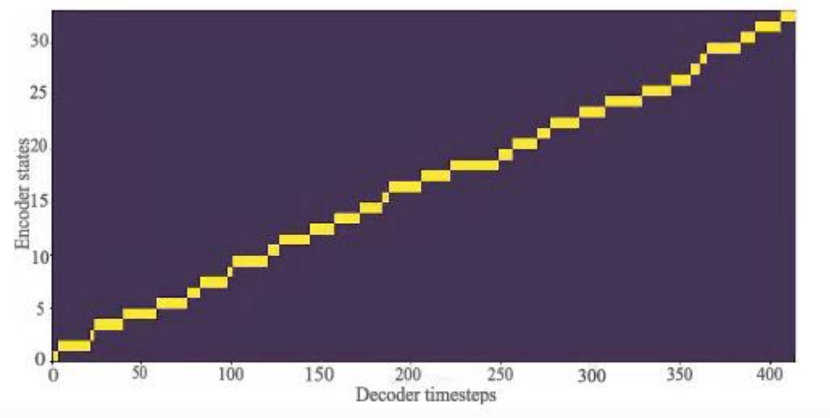
Stop token: 발화의 종료 시점을 추정할 수 있음

WaveNet 보코더: 합성음 품질을 더 더 더욱 높임

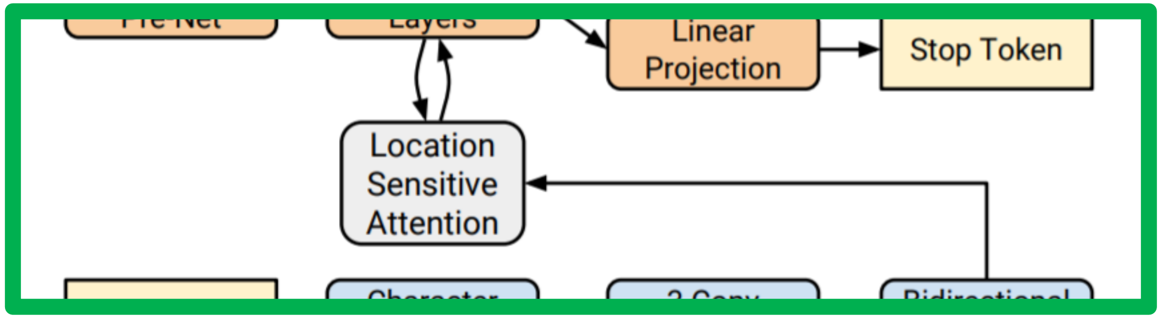
Neural TTS 패러다임을 이끌어낸 주인공 (?)

Acoustic model

Tacotron 2



Alignment



Acoustic model

Tacotron 2

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

End-to-end acoustic model + WaveNet vocoder

당시 최고 합성 모델인 Concatenative 보다 우수한, 녹음에 가까운 수준의 음성 합성 모델

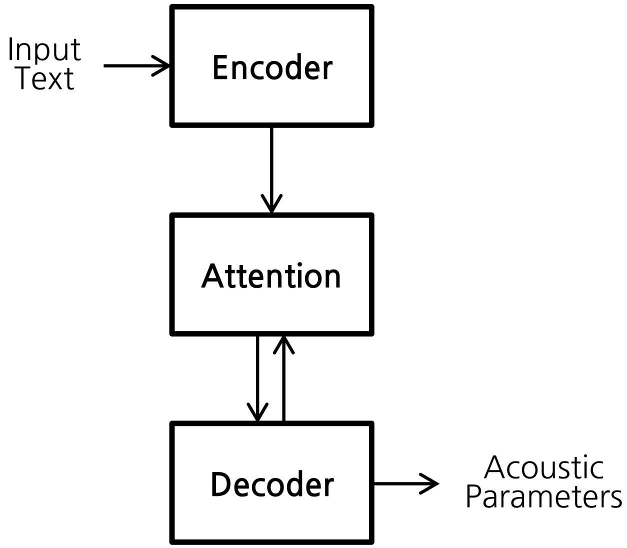
https://ai.googleblog.com/2017/12/tacotron-2-generating-human-like-speech.html

Acoustic model

Summary

End-to-end speech synthesis

- Seq2seq model



Tacotron 2

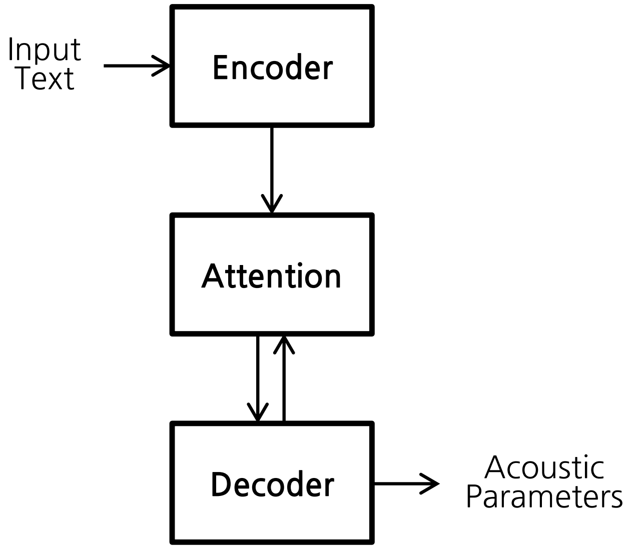
Seq2seq + attention	**Feature engineering 최소화**
Autoregressive decoder	**품질 향상**
Neural vocoder (WaveNet)	품질 향상

Acoustic model

Summary

End-to-end speech synthesis

- Seq2seq model



Tacotron 2

Seq2seq + attention **Feature engineering 최소화**

Autoregressive decoder **품질 향상**

Neural vocoder (WaveNet) 품질 향상

Alignment failure

Slow inference

Acoustic model

Non-autoregressive TTS: FastSpeech 2

FASTSPEECH 2: FAST AND HIGH-QUALITY END-TO-END TEXT TO SPEECH

Yi Ren[1]*, Chenxu Hu[1]*, Xu Tan[2], Tao Qin[2], Sheng Zhao[3], Zhou Zhao[1]†, Tie-Yan Liu[2]

[1]Zhejiang University
{rayeren,chenxuhu,zhaozhou}@zju.edu.cn

[2]Microsoft Research Asia
{xuta,taoqin,tyliu}@microsoft.com

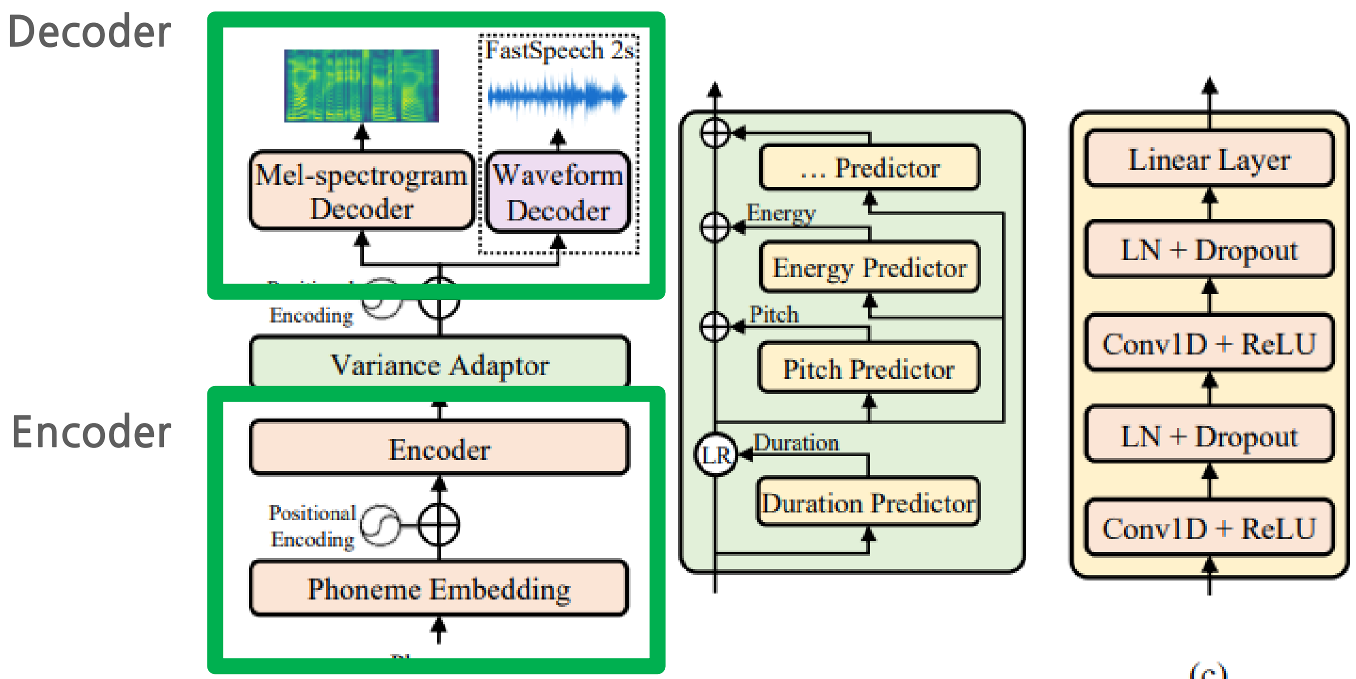
[3]Microsoft Azure Speech
Sheng.Zhao@microsoft.com

ABSTRACT

Non-autoregressive text to speech (TTS) models such as FastSpeech (Ren et al., 2019) can synthesize speech significantly faster than previous autoregressive models with comparable quality. The training of FastSpeech model relies on an autoregressive teacher model for duration prediction (to provide more information as input) and knowledge distillation (to simplify the data distribution in output), which can ease the one-to-many mapping problem (i.e., multiple speech variations correspond to the same text) in TTS. However, FastSpeech has several disadvantages: 1) the teacher-student distillation pipeline is complicated and time-consuming, 2) the duration extracted from the teacher model is not accurate enough, and the target mel-spectrograms distilled from teacher model suffer from information loss due to data simplification, both of which limit the voice quality. In this paper, we propose FastSpeech 2, which addresses the issues in FastSpeech and better solves the one-to-many mapping problem in TTS by 1) directly training the model with ground-truth target instead of the simplified output from teacher, and 2) introducing more variation information of speech (e.g., pitch, energy and more accurate duration) as conditional inputs. Specifically, we extract duration, pitch and energy from speech waveform and directly take them as conditional inputs in training and use predicted values in inference. We further design FastSpeech 2s, which is the first attempt to directly generate speech waveform from text in parallel, enjoying the benefit of fully end-to-end inference. Experimental results show that 1) FastSpeech 2 achieves a 3x training speed-up over FastSpeech, and FastSpeech 2s enjoys even faster inference speed; 2) FastSpeech 2 and 2s outperform FastSpeech in voice quality, and FastSpeech 2 can even surpass autoregressive models. Audio samples are available at https://speechresearch.github.io/fastspeech2/.

Acoustic model

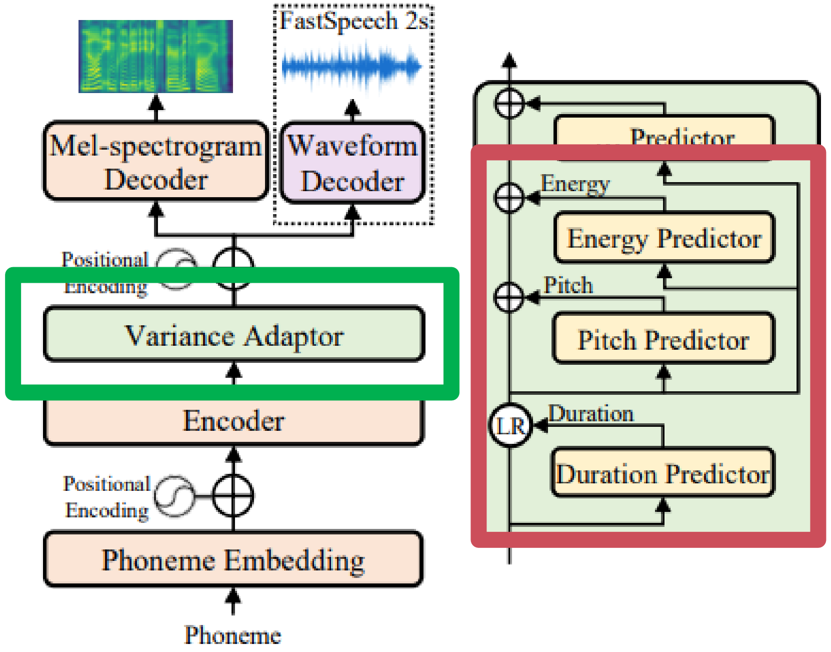
Non-autoregressive TTS: FastSpeech 2



(a) FastSpeech 2 (b) Variance adaptor (c) Duration/pitch/energy predictor

Acoustic model

Non-autoregressive TTS: FastSpeech 2



(a) FastSpeech 2 (b) Variance adaptor

Transformer 기반의 encoder–decoder model

Variance adaptor: Controllability 제공

Duration predictor & Length regulator

Pitch regulator & Energy regulator

좀 비싸도 phoneme segmentation 하고

Knowledge distillation 없이 학습하자!

대신 서비스에 맞게 다양한 기능 넣자!

Acoustic model

Non-autoregressive TTS: FastSpeech 2

Method	MOS
GT	4.30 ± 0.07
GT (Mel + PWG)	3.92 ± 0.08
Tacotron 2 (Shen et al., 2018) (Mel + PWG)	3.70 ± 0.08
Transformer TTS (Li et al., 2019) (Mel + PWG)	3.72 ± 0.07
FastSpeech (Ren et al., 2019) (Mel + PWG)	3.68 ± 0.09
FastSpeech 2 (Mel + PWG)	3.83 ± 0.08
FastSpeech 2s	3.71 ± 0.09

Method	Training Time (h)	Inference Speed (RTF)	Inference Speedup
Transformer TTS (Li et al., 2019)	38.64	9.32×10^{-1}	/
FastSpeech (Ren et al., 2019)	53.12	1.92×10^{-2}	48.5×
FastSpeech 2	**17.02**	1.95×10^{-2}	47.8×
FastSpeech 2s	92.18	$\mathbf{1.80 \times 10^{-2}}$	**51.8×**

V100 GPU 1장 기준

AR model (Tacotron, Transformer) 보다 품질도 좋고

FastSpeech 하고 합성 속도도 비슷하면서

합성음 컨트롤이 가능함

https://speechresearch.github.io/fastspeech2/

Acoustic model

Summary

Estimating acoustic parameters from text inputs

Statistical Parametric Speech Synthesis

가볍고, 빠르고, 안정적 but 품질이 아쉬움

Acoustic model

Summary

Estimating acoustic parameters from text inputs

End-to-end Speech Synthesis

AR models (Tacotron, Transformer) with Attention Alignment

복잡한 Feature Engineering 최소화 하면서도 고품질의 음성을 만들 수 있음

but 느리고 안전성 떨어짐

Acoustic model

Summary

Estimating acoustic parameters from text inputs

End-to-end Speech Synthesis

Non-AR models (FastSpeech 2) with External Duration Model

빠르고 안정적인 합성음을 만들 수 있음

음질은 Best-quality 일까?

Acoustic model

읽어봅시다

#1: Flow-based acoustic model

Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search

Jaehyeon Kim
Kakao Enterprise
jay.xyz@kakaoenterprise.com

Sungwon Kim
Data Science & AI Lab.
Seoul National University
ksw0306@snu.ac.kr

Jungil Kong
Kakao Enterprise
henry.k@kakaoenterprise.com

Sungroh Yoon*
Data Science & AI Lab.
Seoul National University
sryoon@snu.ac.kr

Abstract

Recently, text-to-speech (TTS) models such as FastSpeech and ParaNet have been proposed to generate mel-spectrograms from text in parallel. Despite the advantage, the parallel TTS models cannot be trained without guidance from autoregressive TTS models as their external aligners. In this work, we propose Glow-TTS, a flow-based generative model for parallel TTS that does not require any external aligner. By combining the properties of flows and dynamic programming, the proposed model searches for the most probable monotonic alignment between text and the latent representation of speech on its own. We demonstrate that enforcing hard monotonic alignments enables robust TTS, which generalizes to long utterances, and employing generative flows enables fast, diverse, and controllable speech synthesis. Glow-TTS obtains an order-of-magnitude speed-up over the autoregressive model, Tacotron 2, at synthesis with comparable speech quality. We further show that our model can be easily extended to a multi-speaker setting.

https://jaywalnut310.github.io/glow-tts-demo/index.html

Acoustic model

읽어봅시다

#2: Diffusion-based acoustic model

Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech

Vadim Popov [*1] **Ivan Vovk** [*12] **Vladimir Gogoryan** [12] **Tasnima Sadekova** [1] **Mikhail Kudinov** [1]

[*]Equal contribution [1]Huawei Noah's Ark Lab, Moscow, Russia [2]Higher School of Economics, Moscow, Russia. Correspondence to: Vadim Popov <vadim.popov@huawei.com>, Ivan Vovk <vovk.ivan@huawei.com>.

Abstract

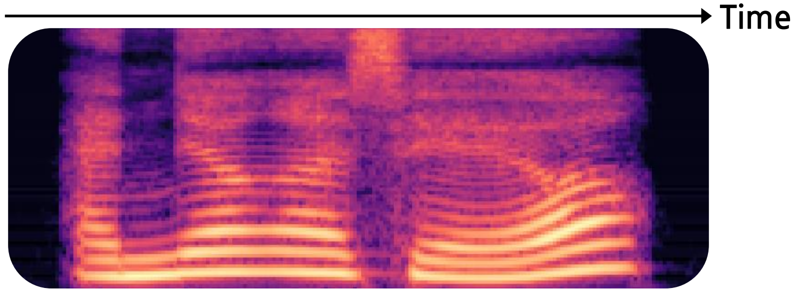
Recently, denoising diffusion probabilistic models and generative score matching have shown high potential in modelling complex data distributions while stochastic calculus has provided a unified point of view on these techniques allowing for flexible inference schemes. In this paper we introduce Grad-TTS, a novel text-to-speech model with score-based decoder producing mel-spectrograms by gradually transforming noise predicted by encoder and aligned with text input by means of Monotonic Alignment Search. The framework of stochastic differential equations helps us to generalize conventional diffusion probabilistic models to the case of reconstructing data from noise with different parameters and allows to make this reconstruction flexible by explicitly controlling trade-off between sound quality and inference speed. Subjective human evaluation shows that Grad-TTS is competitive with state-of-the-art text-to-speech approaches in terms of Mean Opinion Score.

https://grad-tts.github.io/

Speech synthesis and its applications

Vocoder

Generating **speech signals** from **acoustic parameters**

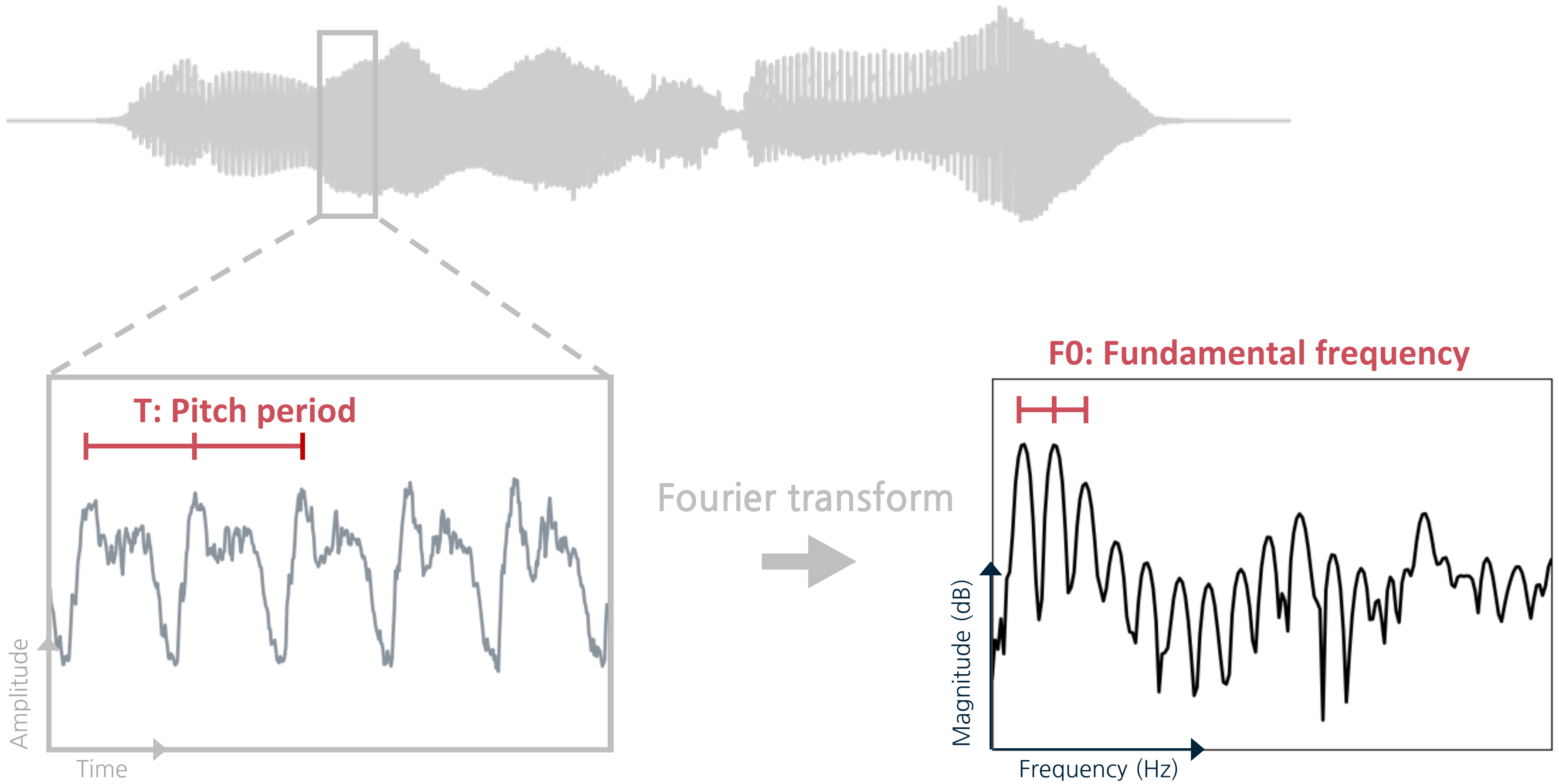


Text → Acoustic Model → Acoustic Parameters → Vocoding Model →

Estimating speech signals from acoustic parameters

How do we produce speech?

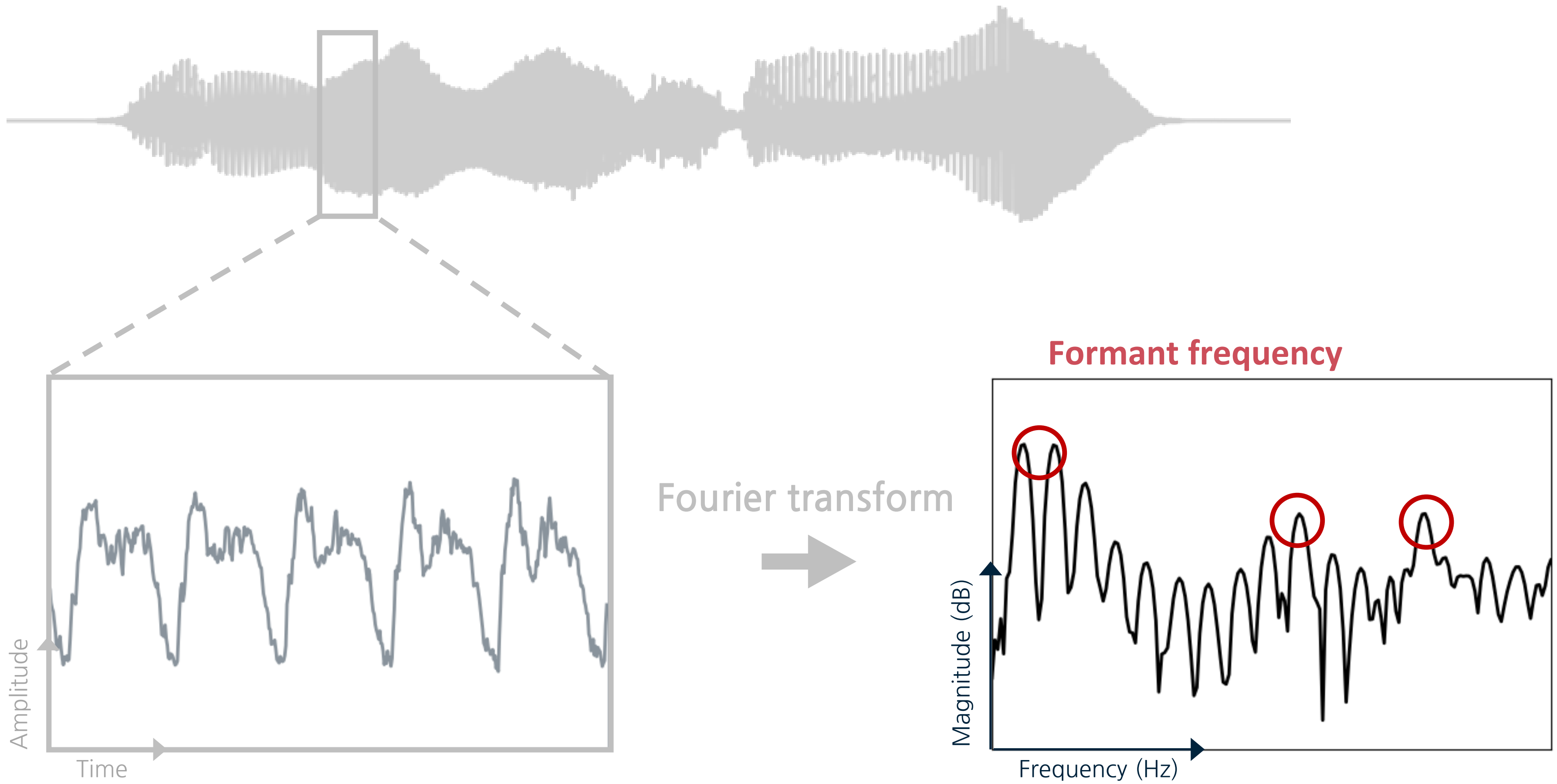
Recall: Speech waveform



F0 의 높낮이에 따라 목소리의 톤이 결정됩니다 (아↘아↗)

How do we produce speech?

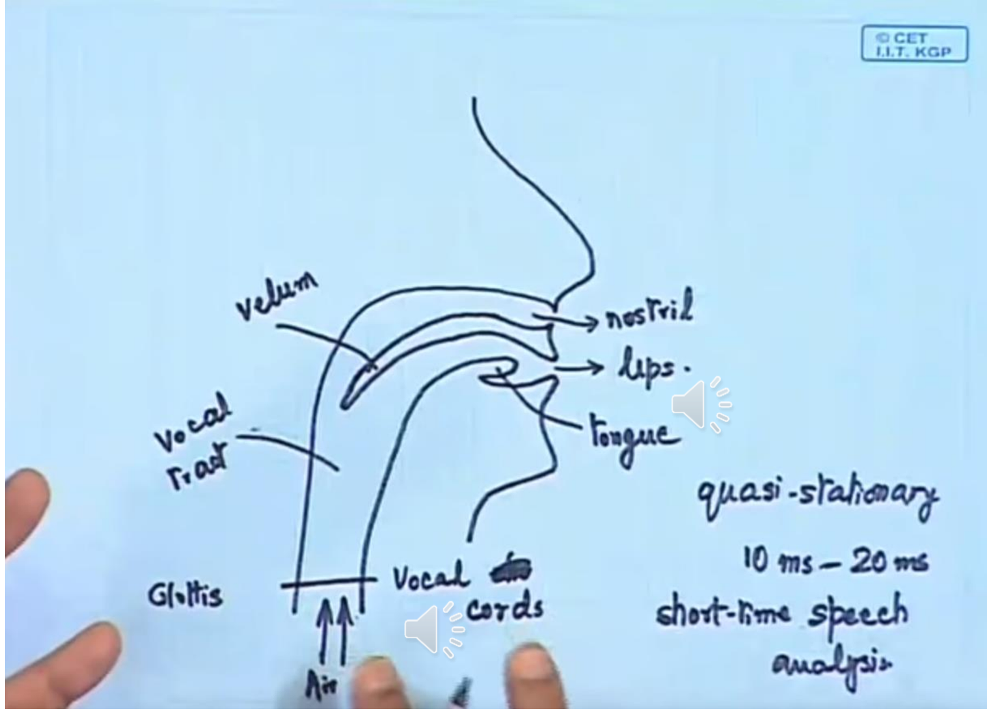
Recall: Speech waveform



Formant frequency 위치에 따라 발음이 결정됩니다 (아/에/이/오/우)

How do we produce speech?

Speech production model



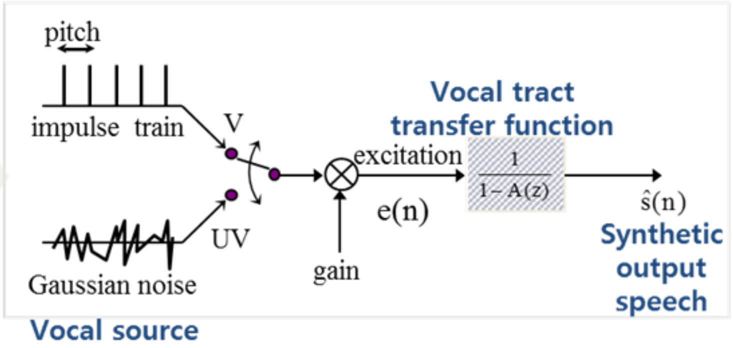
https://www.youtube.com/watch?v=X_JvfZiGEek

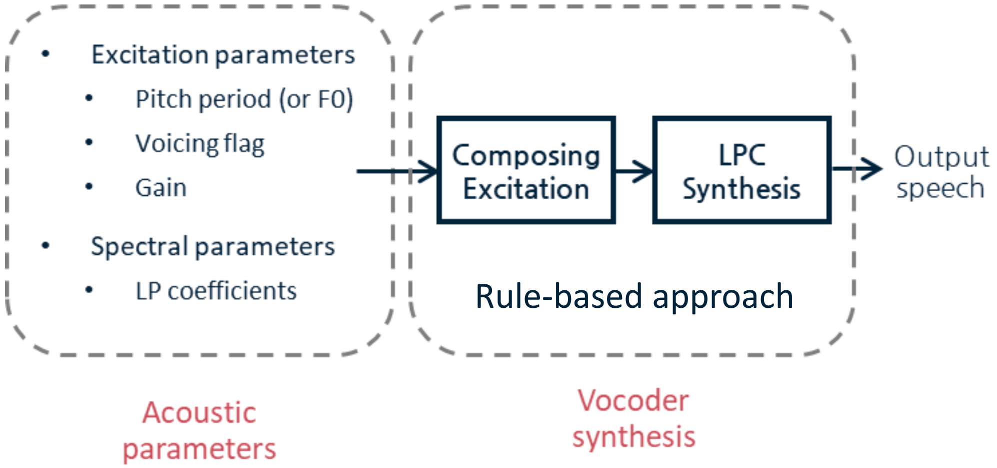
- Lung
  - Power supply
- Vocal source
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract filter
  - Shaping voice color

Source → $\boxed{\textbf{Filter}}$ → **Speech**

Vocoder = Voice + Coder

Parametric approach

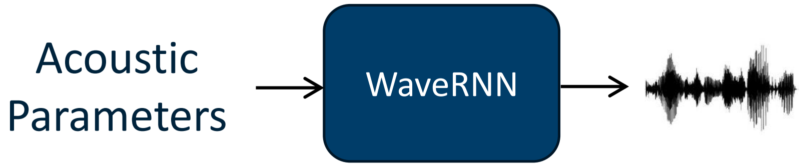




Limitations ☹
- Feature engineering
- Synthetic quality

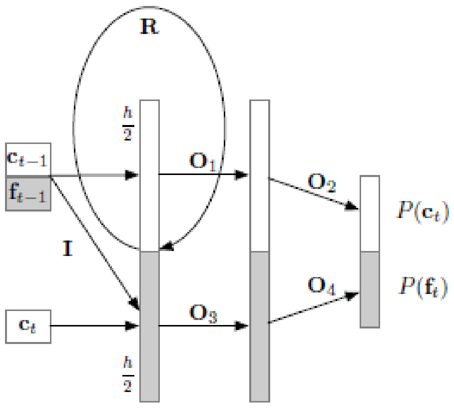
Neural vocoder

Generating speech signals from acoustic parameters



What is the main model?

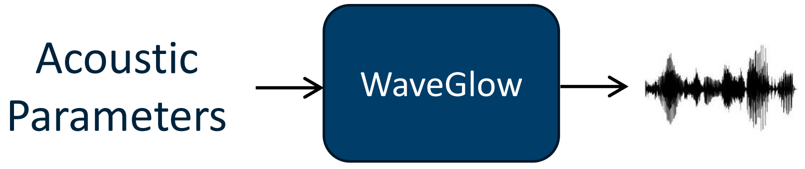
WaveRNN based on the RNN model



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

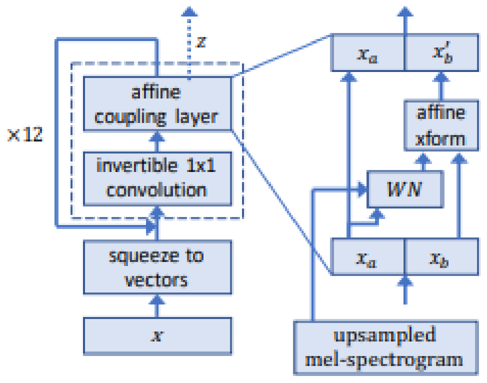
Neural vocoder

Generating 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.

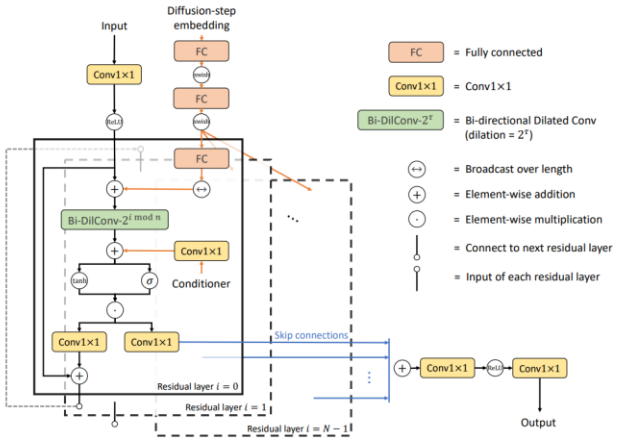
Neural vocoder

Generating **speech signals** from **acoustic parameters**

Acoustic Parameters → DiffWave → 

What is the main model?

DiffWave based on the Diffusion model

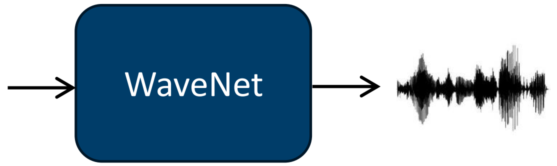


Z. Kong, et al., "Diffwave: A versatile diffusion model for audio synthesis," in Proc. ICLR, 2021.

Neural vocoder

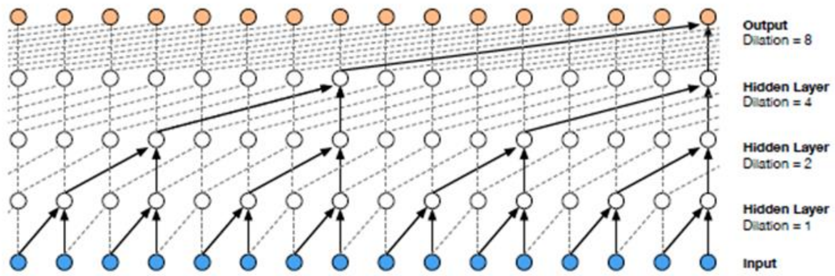
WaveNet synthesis

Acoustic Parameters



What is the main model?

WaveNet based on the CNN model



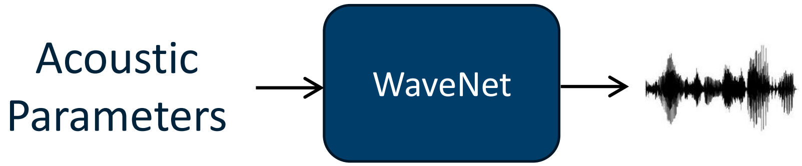
$$p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^{T} p(x_t|x_1, \cdots, x_{t-1}, \mathbf{h})$$

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

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

Neural vocoder

WaveNet synthesis



What is the main model?

WaveNet based on the CNN model

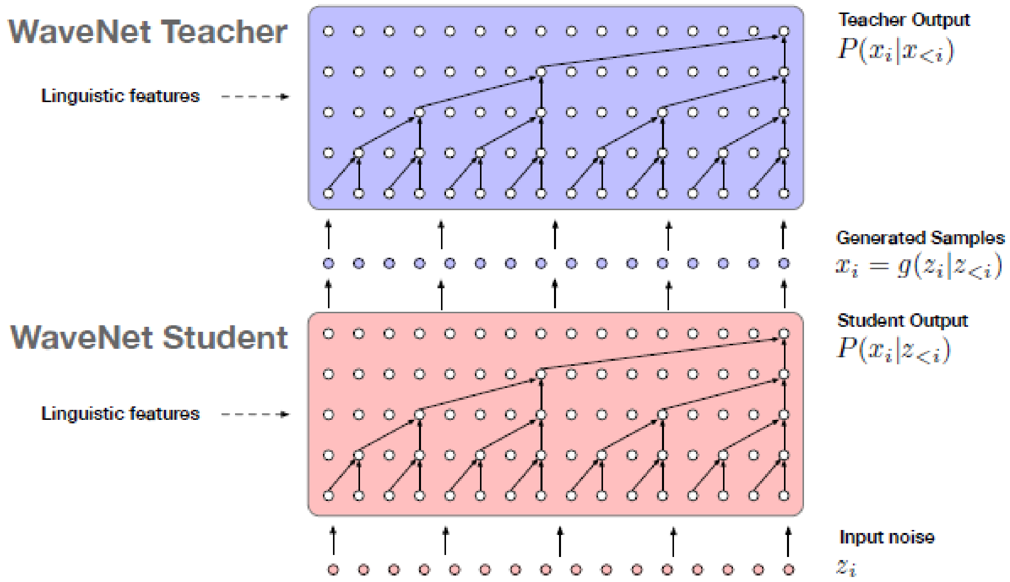
Estimating the current sample from the previous samples
We define this method as autoregressive vocoding model

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.

Neural vocoder

Parallel WaveNet synthesis

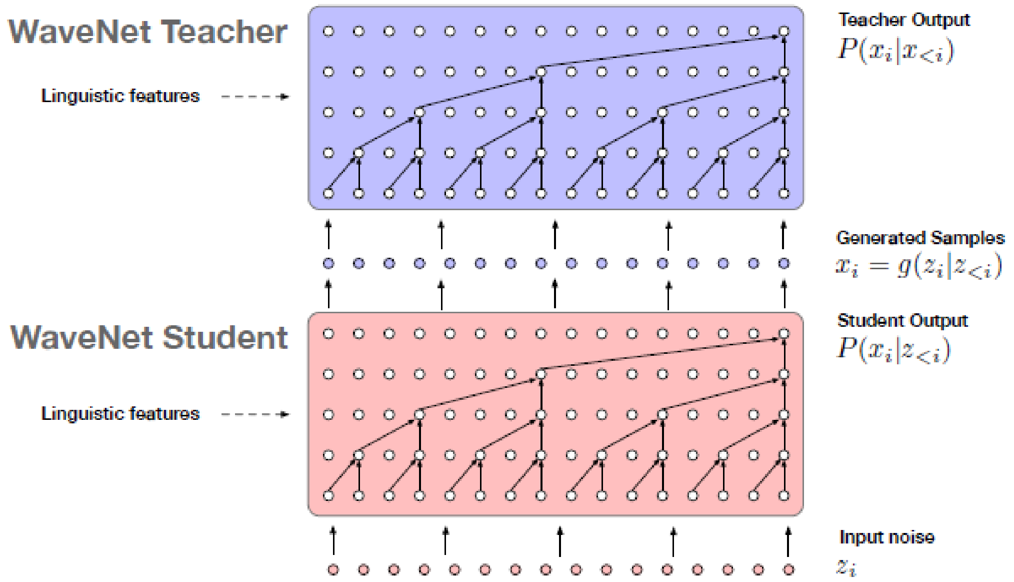


One of the alternative method to address WaveNet's slow inference speed is the non-autoregressive Parallel WaveNet

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

Neural vocoder

Parallel WaveNet synthesis

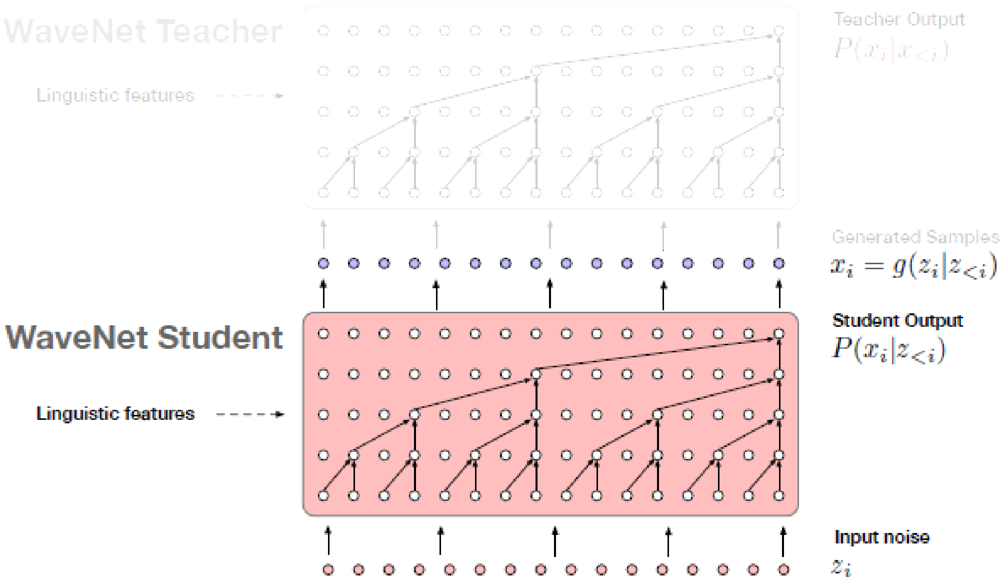


Non-autoregressive Parallel WaveNet (=student) is trained to learn the distribution of the autoregressive WaveNet (=teachure)

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

Neural vocoder

Parallel WaveNet synthesis

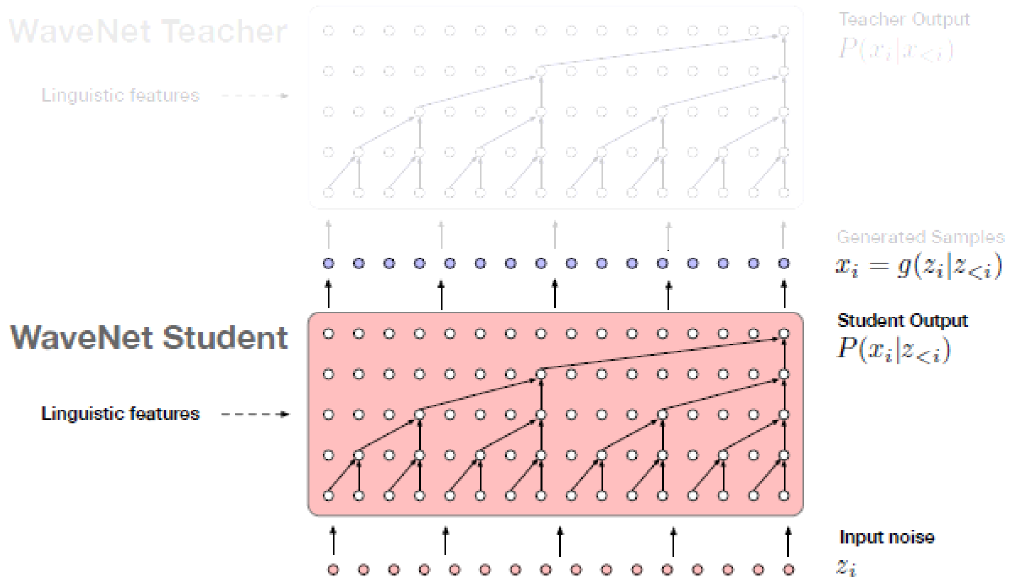


Non-autoregressive Parallel WaveNet doesn't require the previous samples
Its inference speed in unlimited
(it takes about 0.02 sec to generate 1 sec audio)

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

Neural vocoder

Parallel WaveNet synthesis



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.

Neural vocoder

Parallel WaveNet synthesis

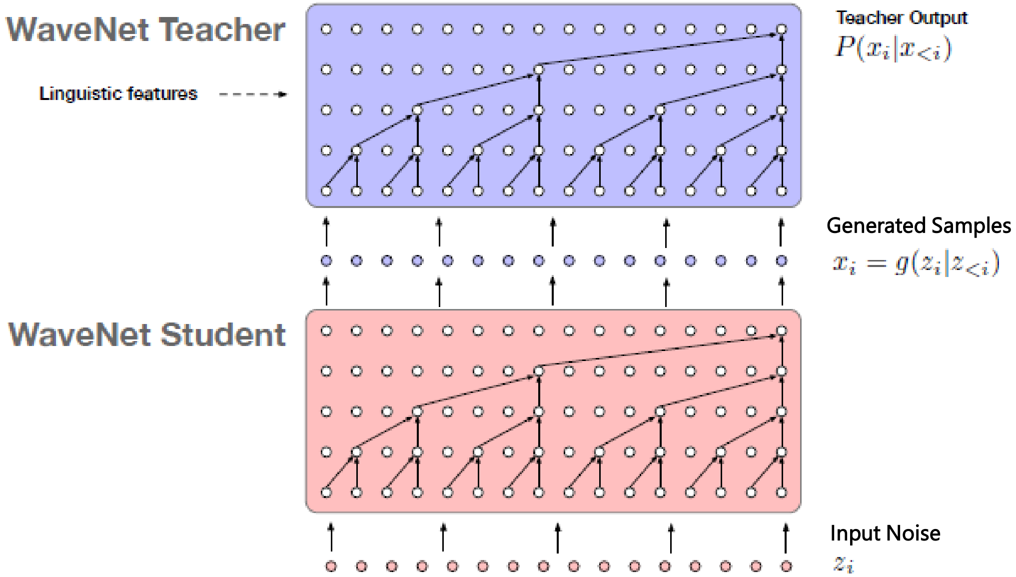
PARALLEL WAVEGAN: A FAST WAVEFORM GENERATION MODEL BASED ON GENERATIVE ADVERSARIAL NETWORKS WITH MULTI-RESOLUTION SPECTROGRAM

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.

Neural vocoder: Parallel WaveGAN

1. Removed the teacher-student distillation process

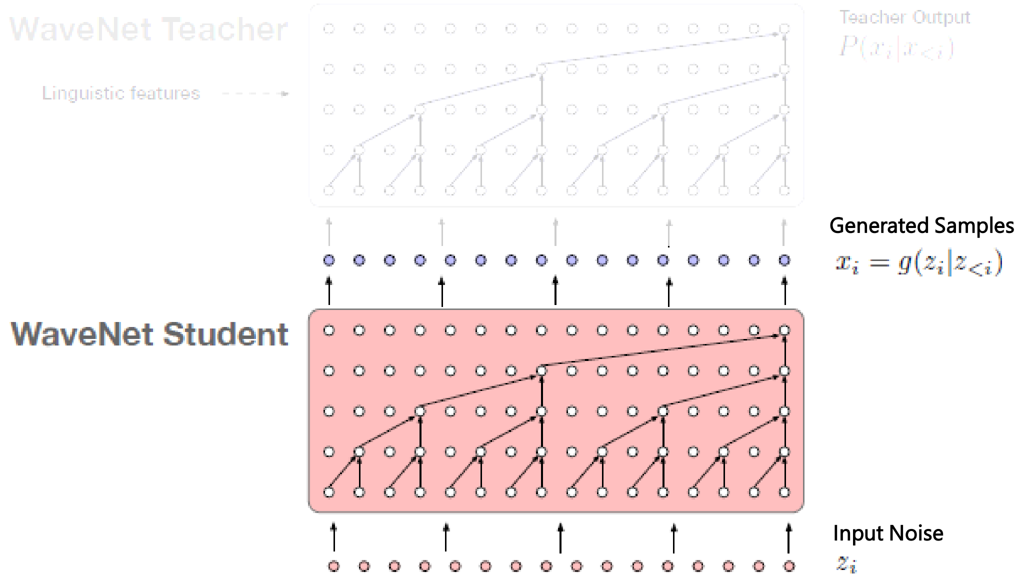


R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

Neural vocoder: Parallel WaveGAN

1. Removed the teacher-student distillation process

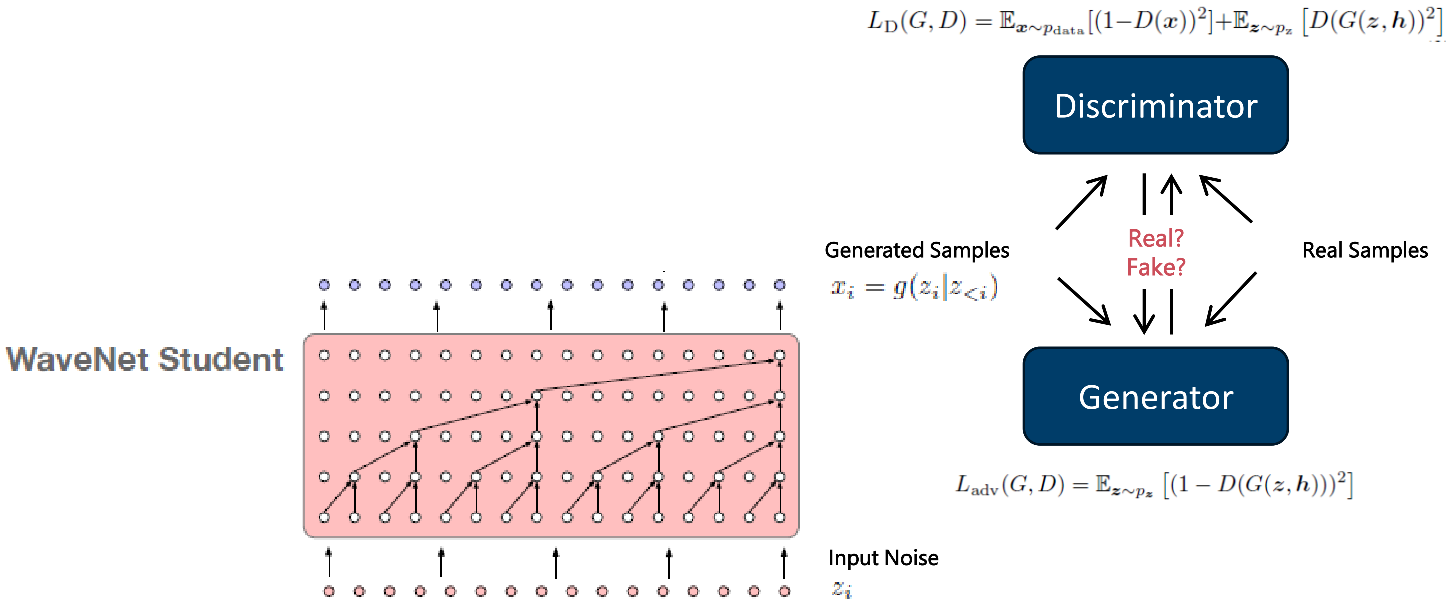
   → Entire model can be "easily" trained



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.

Neural vocoder: Parallel WaveGAN

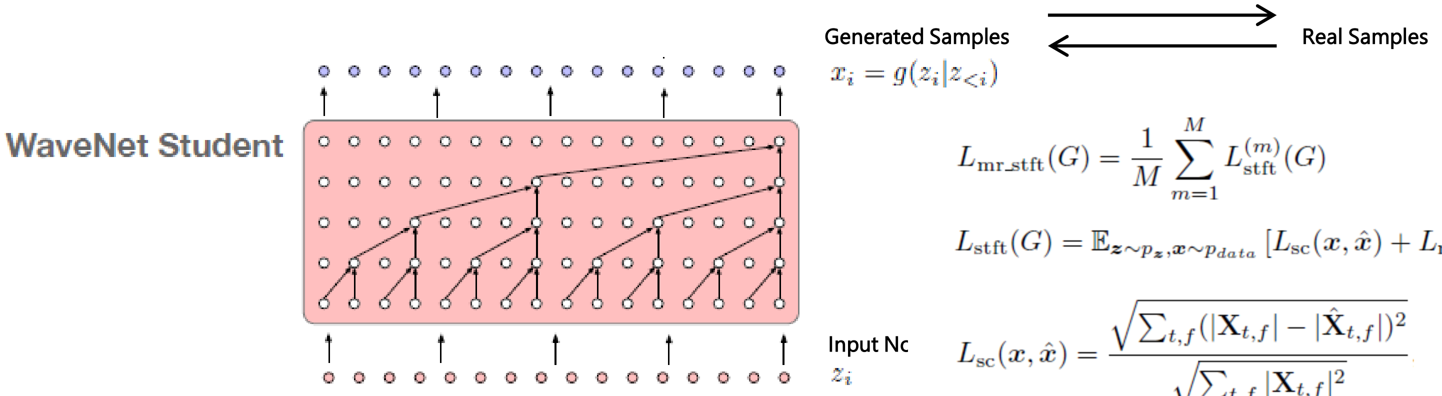
1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method



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.

Neural vocoder: 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



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

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

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

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

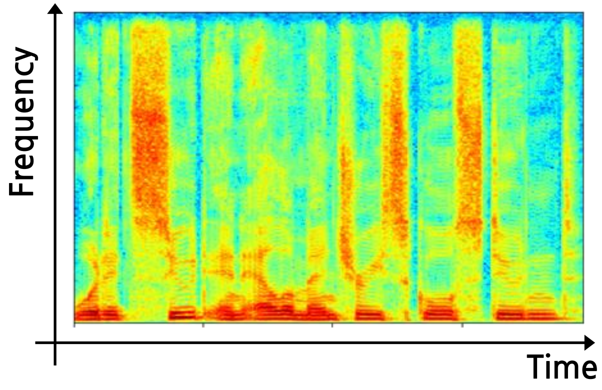
R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

Neural vocoder: 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

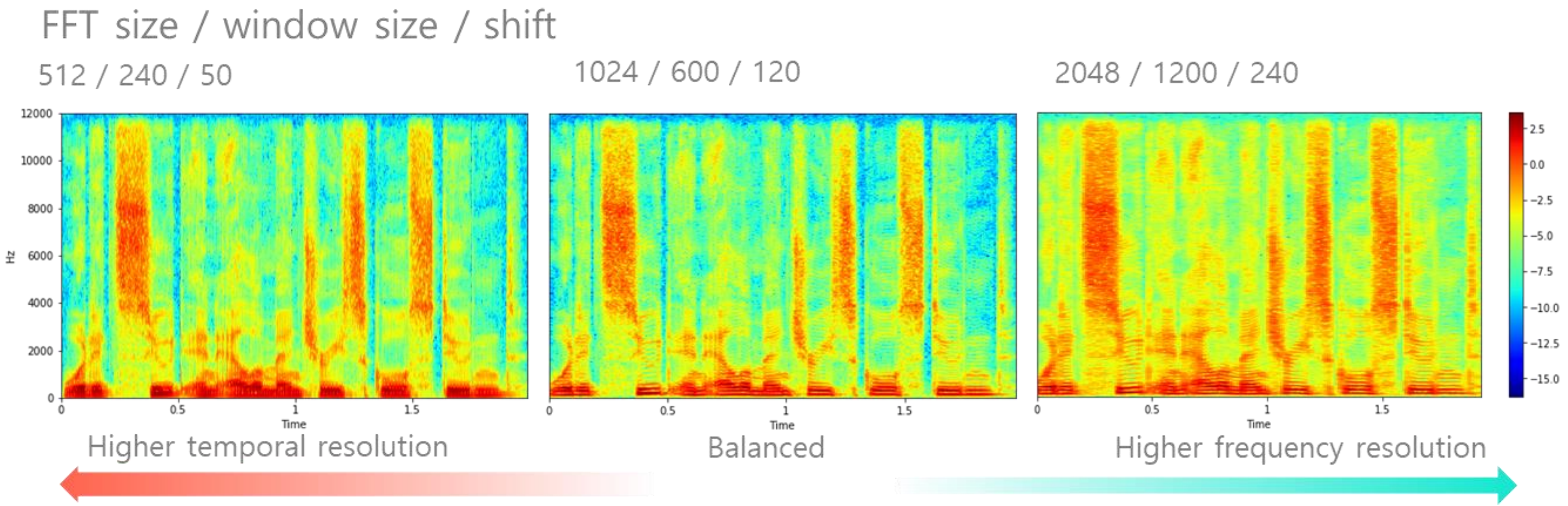


R. Yamamoto, et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6194-6198.

Neural vocoder: 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



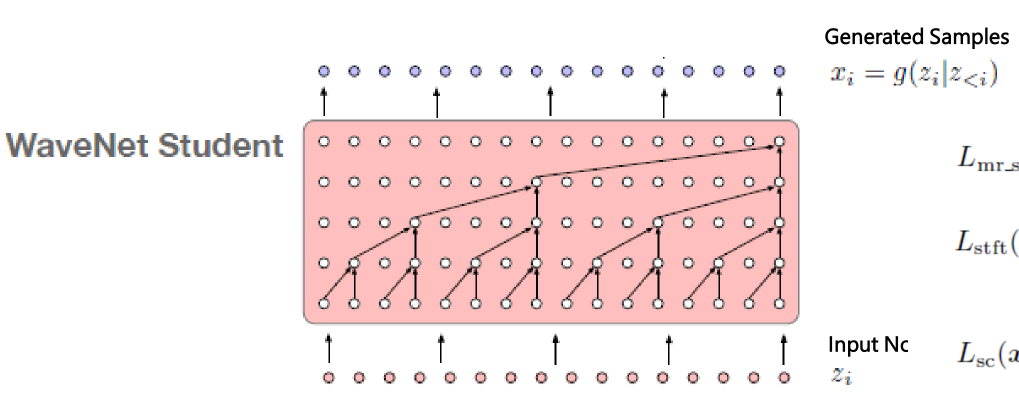
R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. ICASSP, 2020, pp. 6194-6198.

Neural vocoder: 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



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

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

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

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

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.

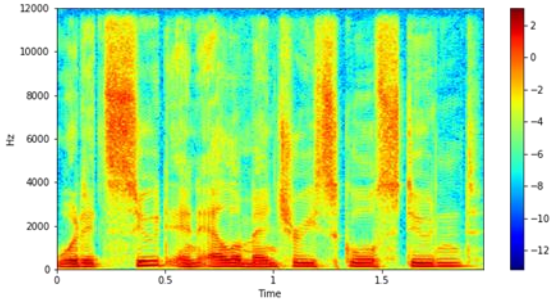
Neural vocoder: Parallel WaveGAN

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

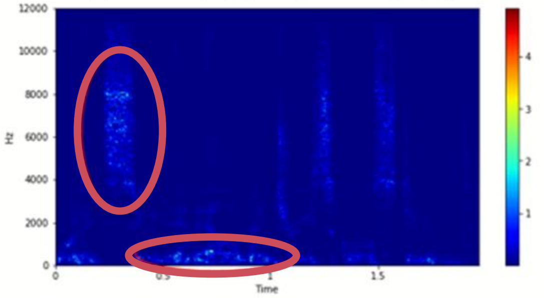
STFT is calculated in different T/F resolutions

There are two loss functions

One penalizes large energy components



Real Spectrogram



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.

Neural vocoder: Parallel WaveGAN

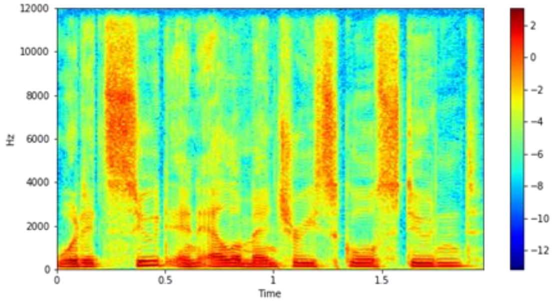
1. Removed the teacher-student distillation process
2. Improved synthetic quality by using the adversarial training method
3. Further improved its quality by introducing the multi-resolution STFT loss

STFT is calculated in different T/F resolutions

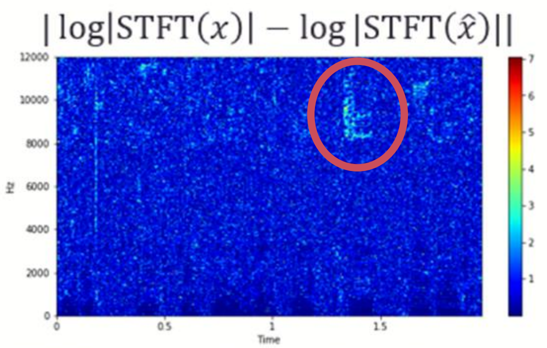
There are two loss functions

One penalizes large energy components

The other penalizes small energy components



Real Spectrogram



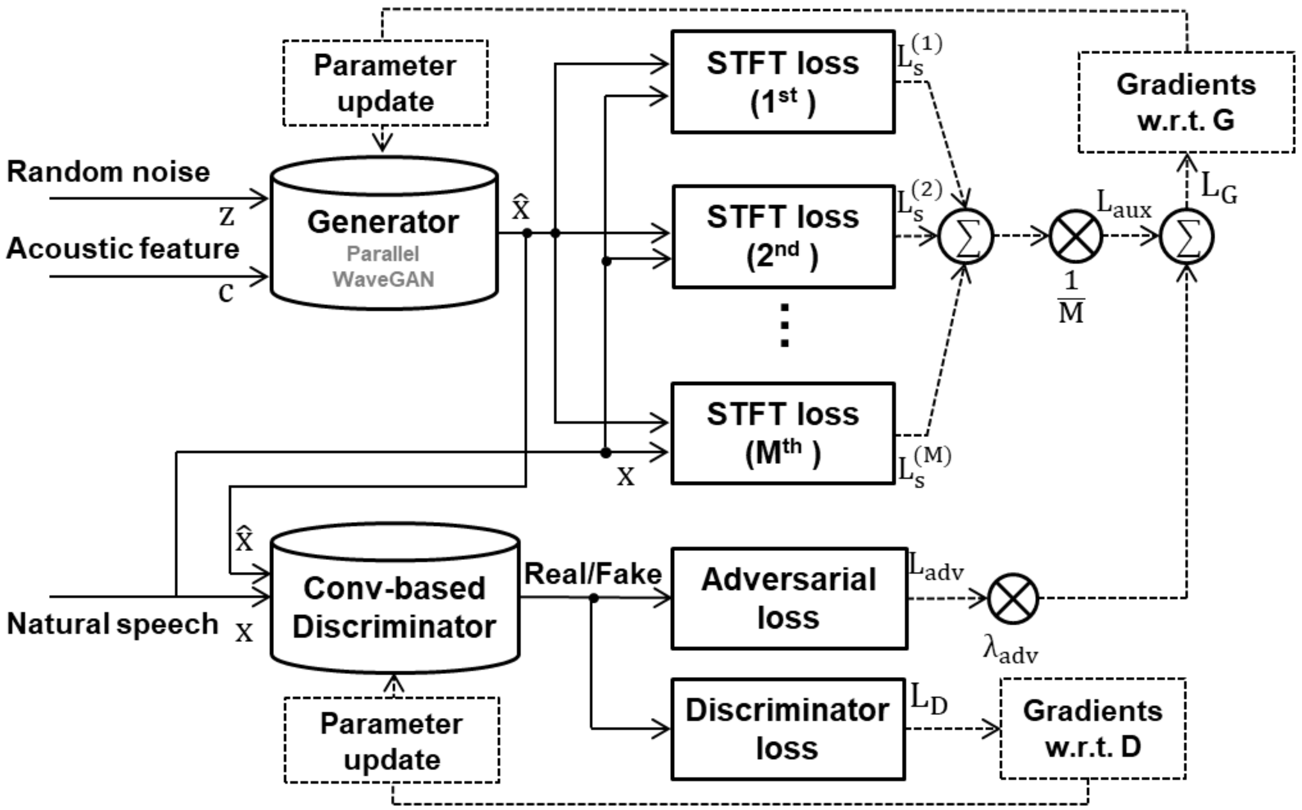
Log STFT loss penalizes small amplitude components

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

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.

Neural vocoder: 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.

Neural vocoder: Parallel WaveGAN

Training method

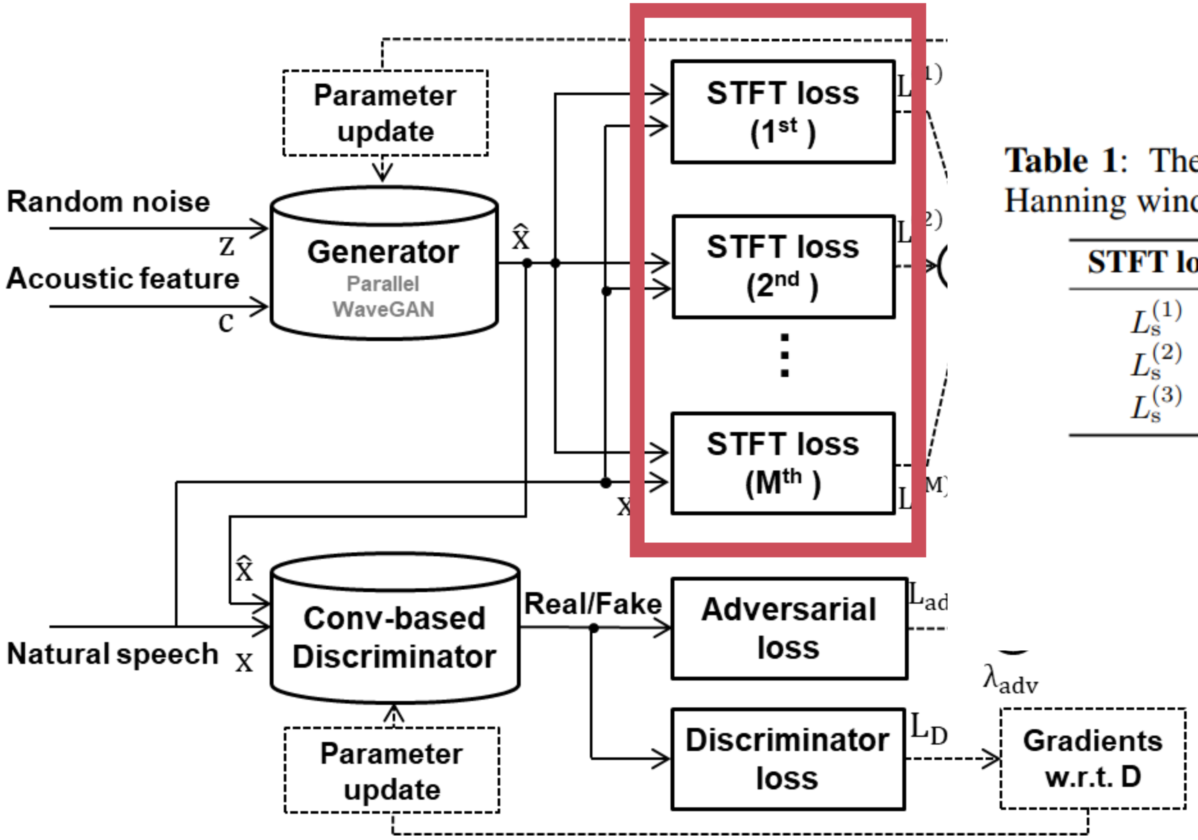


Table 1: The details of the multi-resolution STFT loss. A Hanning window was applied before the FFT process.

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 ($\approx$ 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.

Neural vocoder: Parallel WaveGAN

Parallel WaveNet synthesis

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

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.

Neural vocoder: 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	0.32×10^{-2}	3.61 ± 0.12
System 2	ClariNet	Yes	$L_s^{(1)}$	-	60	2.78 M	14.62	3.88 ± 0.11
System 3	ClariNet	Yes	$L_s^{(1)}+L_s^{(2)}+L_s^{(3)}$	-	60	2.78 M	14.62	4.21 ± 0.09
System 4	ClariNet	Yes	$L_s^{(1)}+L_s^{(2)}+L_s^{(3)}$	Yes	60	2.78 M	14.62	4.21 ± 0.09
System 5	Parallel WaveGAN	-	$L_s^{(1)}$	Yes	30	1.44 M	28.68	1.36 ± 0.07
System 6	Parallel WaveGAN	-	$L_s^{(1)}+L_s^{(2)}+L_s^{(3)}$	Yes	30	1.44 M	28.68	4.06 ± 0.10
System 7	Recording	-	-	-	-			4.46 ± 0.08

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 ± 0.11
Transformer + ClariNet	4.00 ± 0.10
Transformer + ClariNet-GAN	4.14 ± 0.10
Transformer + Parallel WaveGAN (ours)	4.16 ± 0.09
Recording	4.46 ± 0.08

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.

Vocoder

Summary

Estimating speech signals from acoustic parameters

Rule-based parametric vocoders

가볍고, 빠르고, 안정적 but 품질이 아쉬움

Vocoder

Summary

Text → Acoustic Model → Acoustic Parameters → Vocoding Model → 

Estimating speech signals from acoustic parameters

Autoregressive (AR) neural vocoder

Domain-specific feature engineering 최소화 하면서도 고품질의 음성을 만들 수 있음

but 생성 속도가 느려도 너무 느림

Vocoder

Summary

Estimating speech signals from acoustic parameters

Non-AR neural vocoder

Teacher-student paradigm 도입으로 보코더의 속도 이슈 해결

but 학습 과정이 복잡하고 합성음 품질이 아쉬워짐

Vocoder

Summary

Estimating speech signals from acoustic parameters

Non-AR neural vocoder

Adversarial training 도입으로 속도 이슈와 학습 이슈를 모두 해결

Neural vocoder

읽어봅시다

#1: HiFi-GAN

HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis

Jungil Kong
Kakao Enterprise
henry.k@kakaoenterprise.com

Jaehyeon Kim
Kakao Enterprise
jay.xyz@kakaoenterprise.com

Jaekyoung Bae
Kakao Enterprise
storm.b@kakaoenterprise.com

Abstract

Several recent work on speech synthesis have employed generative adversarial networks (GANs) to produce raw waveforms. Although such methods improve the sampling efficiency and memory usage, their sample quality has not yet reached that of autoregressive and flow-based generative models. In this work, we propose HiFi-GAN, which achieves both efficient and high-fidelity speech synthesis. As speech audio consists of sinusoidal signals with various periods, we demonstrate that modeling periodic patterns of an audio is crucial for enhancing sample quality. A subjective human evaluation (mean opinion score, MOS) of a single speaker dataset indicates that our proposed method demonstrates similarity to human quality while generating 22.05 kHz high-fidelity audio 167.9 times faster than real-time on a single V100 GPU. We further show the generality of HiFi-GAN to the mel-spectrogram inversion of unseen speakers and end-to-end speech synthesis. Finally, a small footprint version of HiFi-GAN generates samples 13.4 times faster than real-time on CPU with comparable quality to an autoregressive counterpart.

https://github.com/jik876/hifi-gan

Neural vocoder

읽어봅시다

#2: BigVGAN

BIGVGAN: A UNIVERSAL NEURAL VOCODER WITH LARGE-SCALE TRAINING

Sang-gil Lee[1*] **Wei Ping**[2†]
Boris Ginsburg[2] **Bryan Catanzaro**[2] **Sungroh Yoon**[1,3†]

[1] Data Science & AI Lab, Seoul National University (SNU)
[2] NVIDIA
[3] AIIS, ASRI, INMC, ISRC, NSI, and Interdisciplinary Program in AI, SNU

tkdrlf9202@snu.ac.kr wping@nvidia.com
bginsburg@nvidia.com bcatanzaro@nvidia.com sryoon@snu.ac.kr

ABSTRACT

Despite recent progress in generative adversarial network (GAN)-based vocoders, where the model generates raw waveform conditioned on acoustic features, it is challenging to synthesize high-fidelity audio for numerous speakers across various recording environments. In this work, we present BigVGAN, a universal vocoder that generalizes well for various out-of-distribution scenarios without fine-tuning. We introduce periodic activation function and anti-aliased representation into the GAN generator, which brings the desired inductive bias for audio synthesis and significantly improves audio quality. In addition, we train our GAN vocoder at the largest scale up to 112M parameters, which is unprecedented in the literature. We identify and address the failure modes in large-scale GAN training for audio, while maintaining high-fidelity output without over-regularization. Our BigVGAN, trained only on clean speech (LibriTTS), achieves the state-of-the-art performance for various zero-shot (out-of-distribution) conditions, including unseen speakers, languages, recording environments, singing voices, music, and instrumental audio. [1] We release our code and model at: https://github.com/NVIDIA/BigVGAN.

https://github.com/NVIDIA/BigVGAN

Speech synthesis and its applications

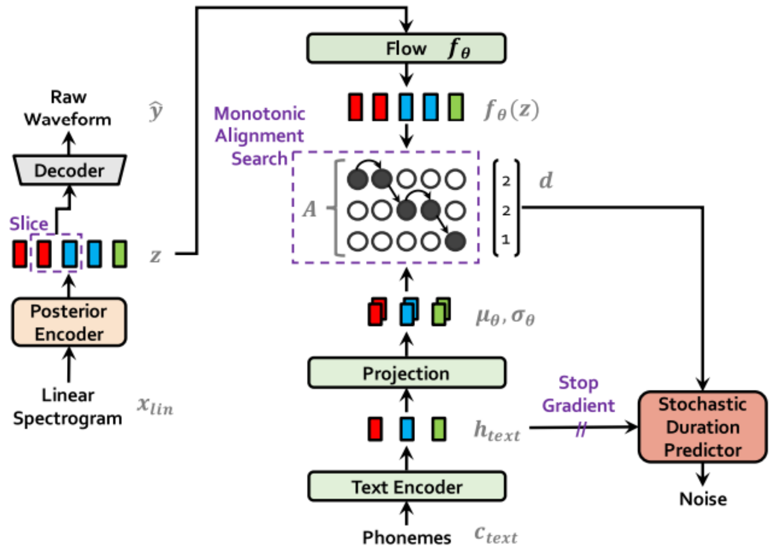
1. Speech analysis: Mel-spectrogram
2. Acoustic models: From text to acoustic parameters
3. Vocoder: From acoustic parameters to speech
4. Fully end-to-end speech synthesis
5. Applications

Fully end-to-end speech synthesis

VITS

Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech

Jaehyeon Kim [1] **Jungil Kong** [1] **Juhee Son** [1 2]



[1]Kakao Enterprise, Seongnam-si, Gyeonggi-do, Republic of Korea [2]School of Computing, KAIST, Daejeon, Republic of Korea. Correspondence to: Jaehyeon Kim <jay.xyz@kakaoenterprise.com>.

Abstract

Several recent end-to-end text-to-speech (TTS) models enabling single-stage training and parallel sampling have been proposed, but their sample quality does not match that of two-stage TTS systems. In this work, we present a parallel end-to-end TTS method that generates more natural sounding audio than current two-stage models. Our method adopts variational inference augmented with normalizing flows and an adversarial training process, which improves the expressive power of generative modeling. We also propose a stochastic duration predictor to synthesize speech with diverse rhythms from input text. With the uncertainty modeling over latent variables and the stochastic duration predictor, our method expresses the natural one-to-many relationship in which a text input can be spoken in multiple ways with different pitches and rhythms. A subjective human evaluation (mean opinion score, or MOS) on the LJ Speech, a single speaker dataset, shows that our method outperforms the best publicly available TTS systems and achieves a MOS comparable to ground truth.

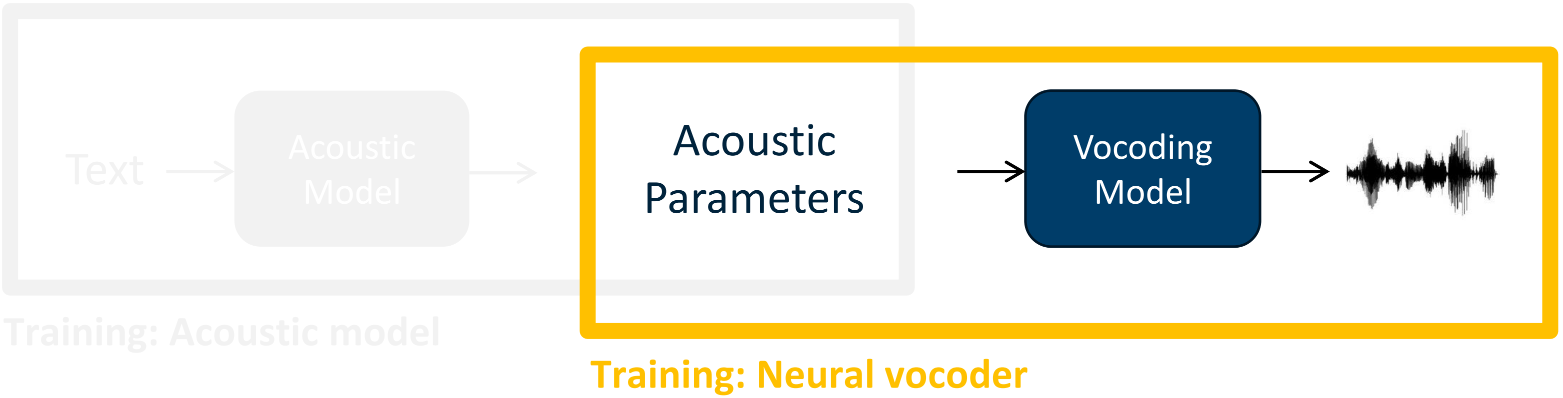
Exposure bias problem

Mismatch between training and inference processes

Training: Acoustic model

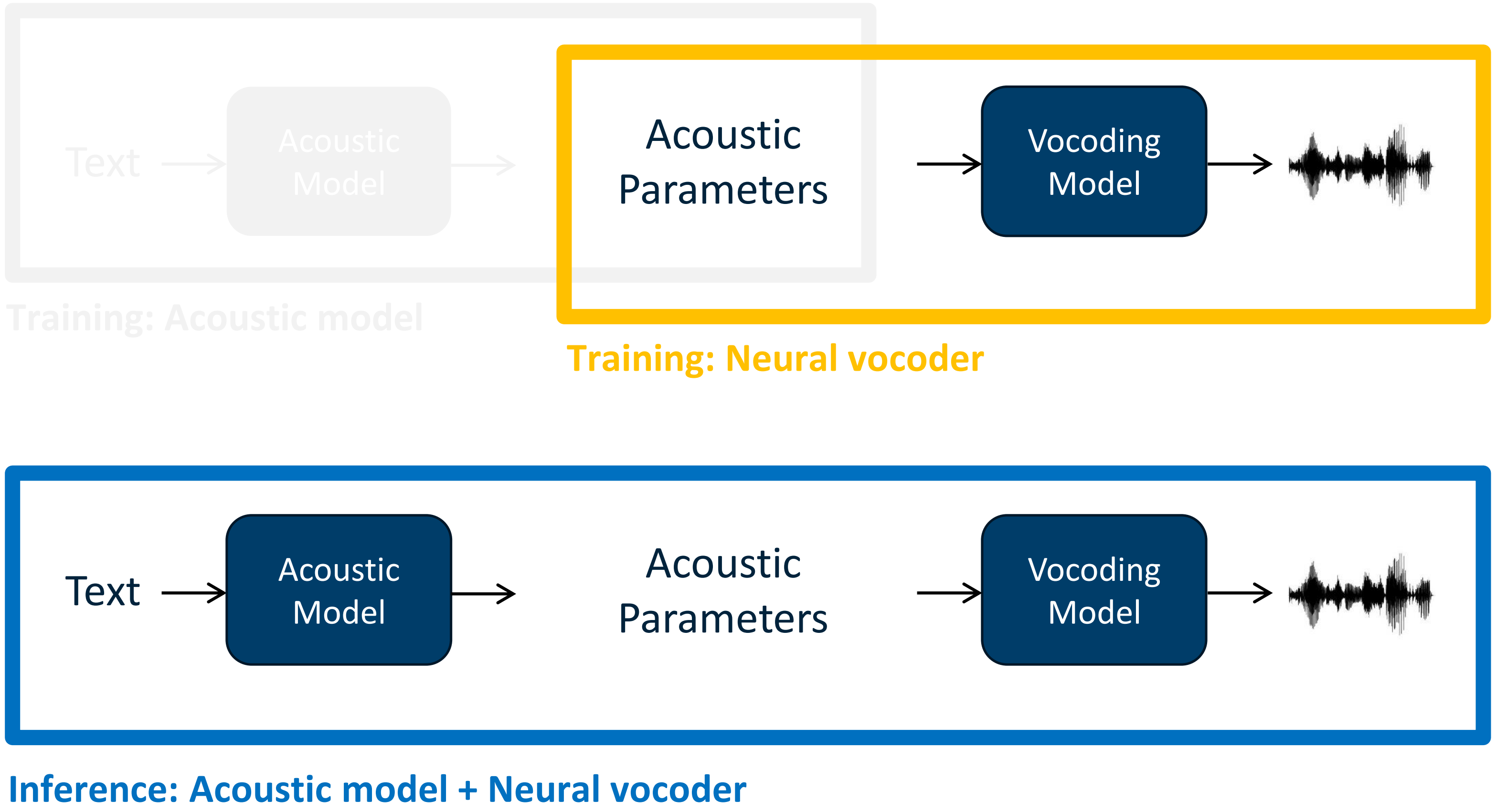
Exposure bias problem

Mismatch between training and inference processes



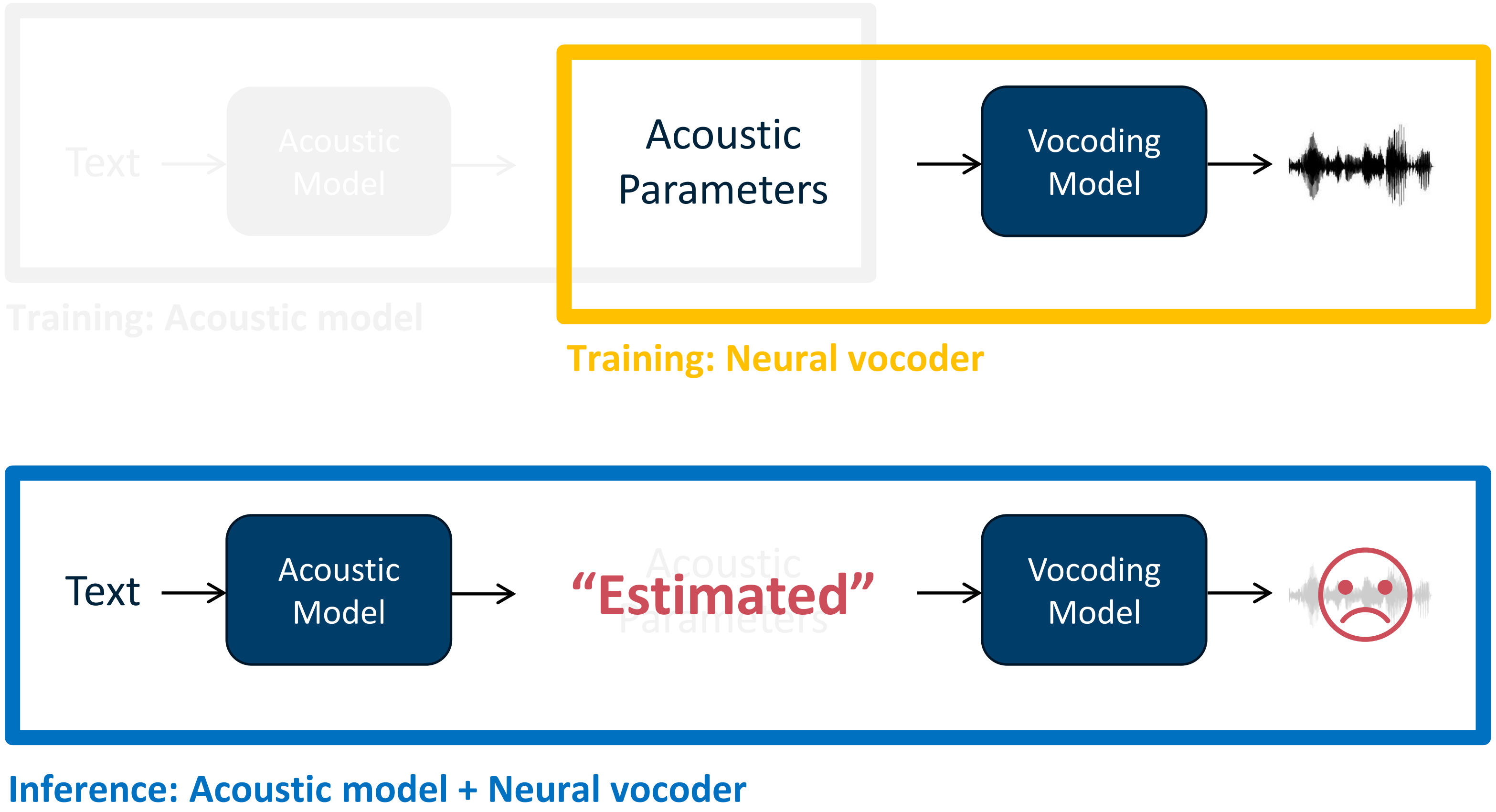
Exposure bias problem

Mismatch between training and inference processes



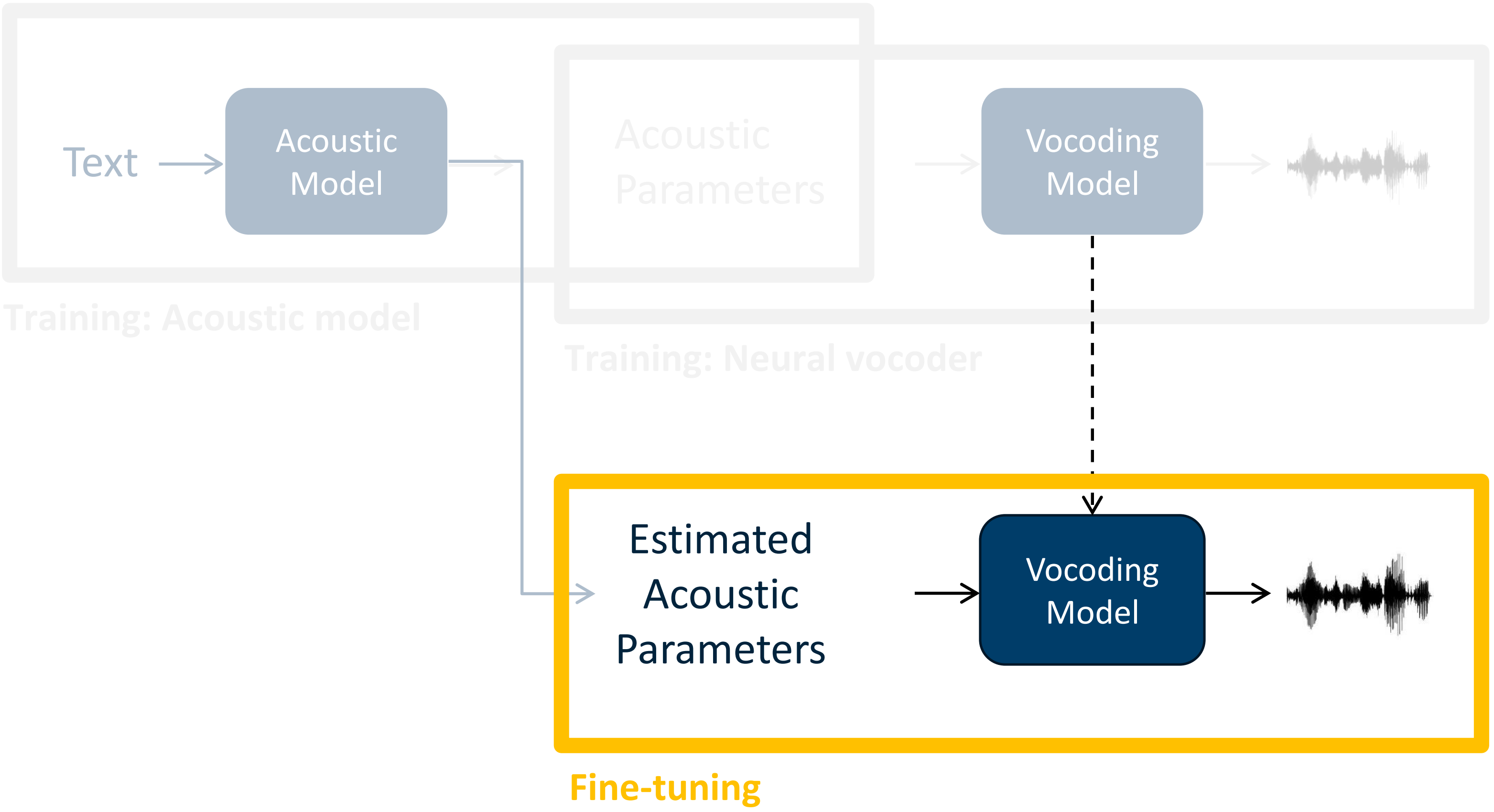
Exposure bias problem

Mismatch between training and inference processes



Exposure bias problem

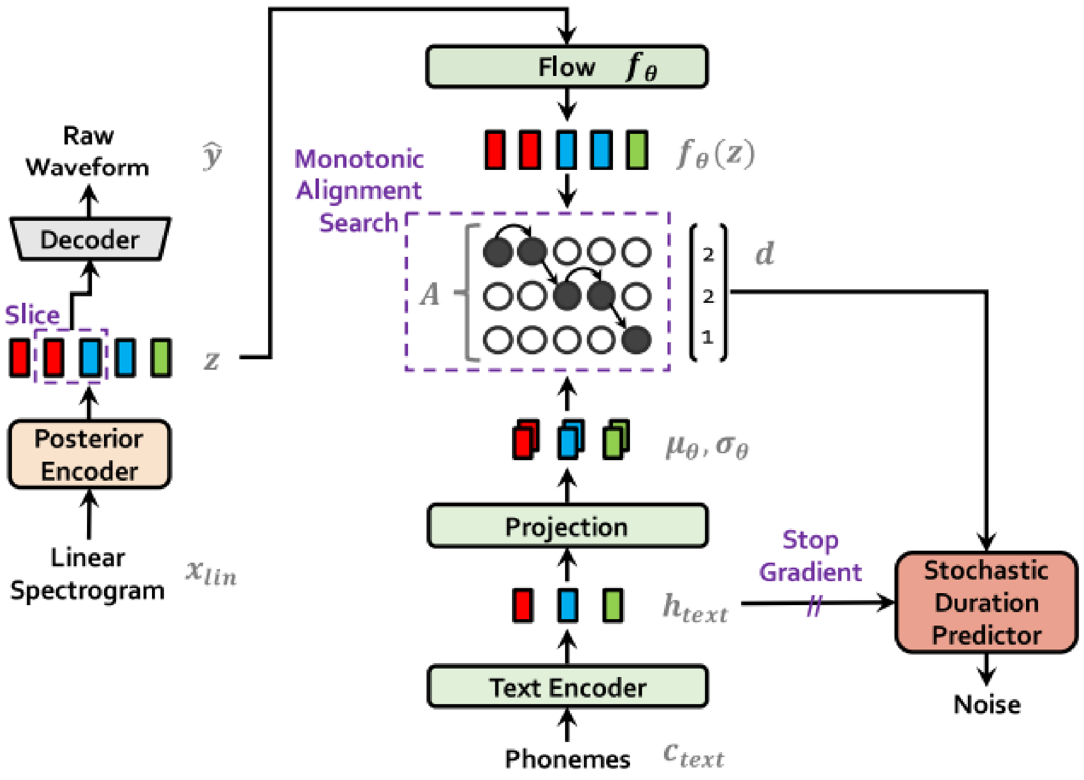
Mismatch between training and inference processes



Inefficient training pipeline ☹

Fully end-to-end speech synthesis

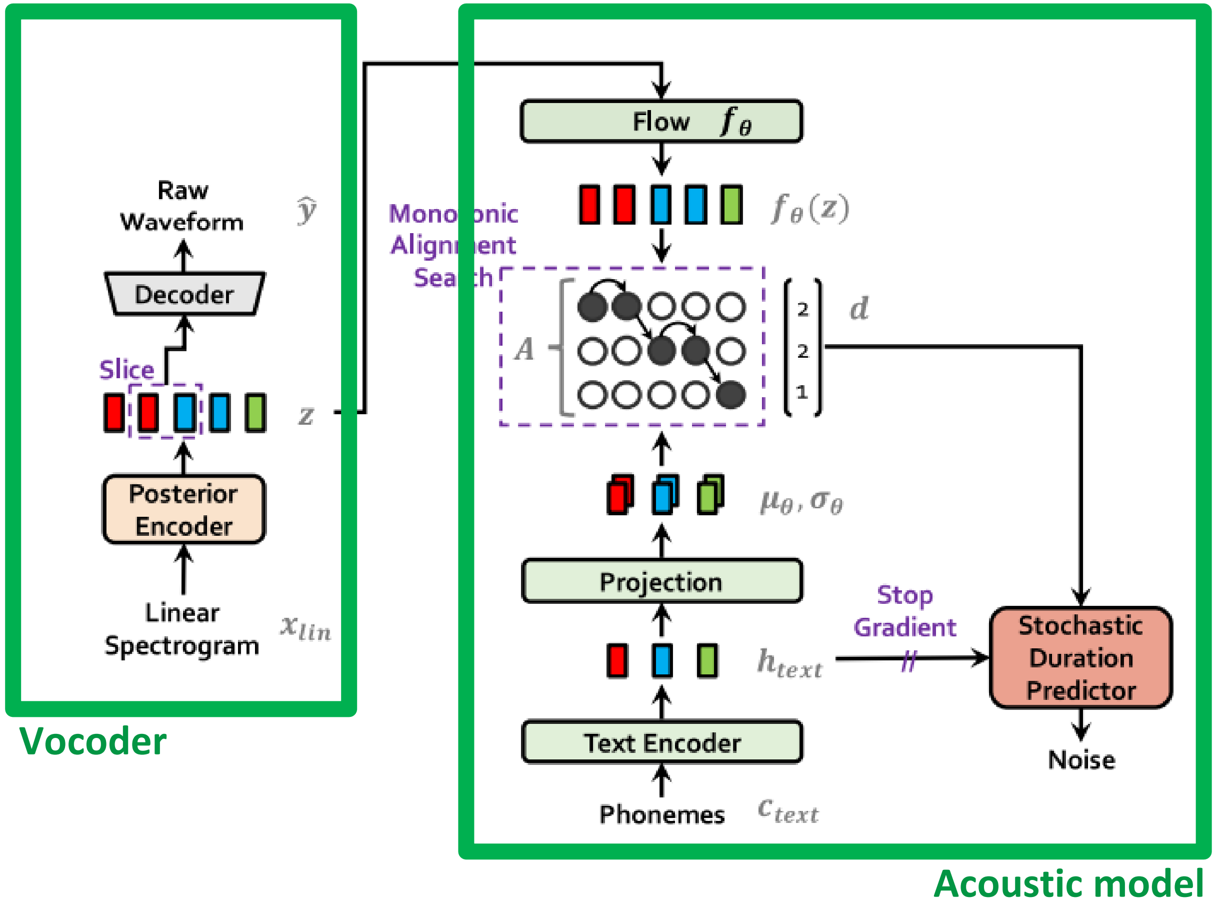
VITS



$$L_{cvae} = L_{recon} + D_{KL}(q(z|y)||p(z|c))$$

Fully end-to-end speech synthesis

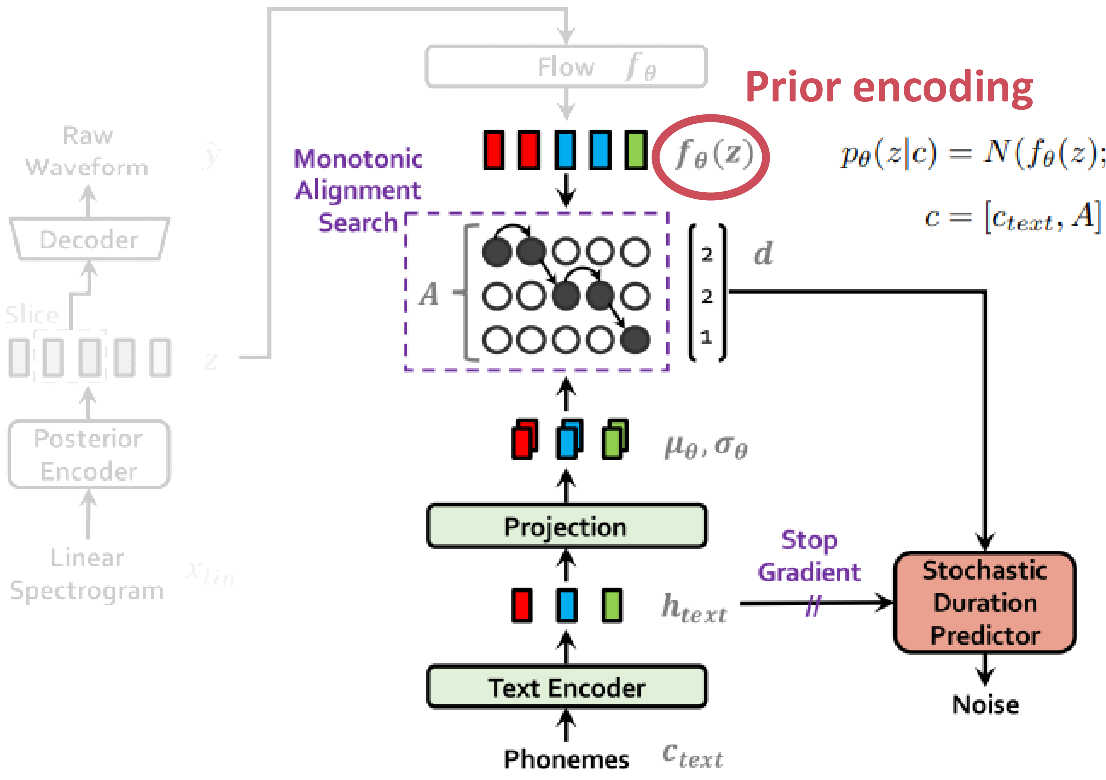
VITS



$$L_{cvae} = L_{recon} + D_{KL}(q(z|y)||p(z|c))$$

Fully end-to-end speech synthesis

VITS



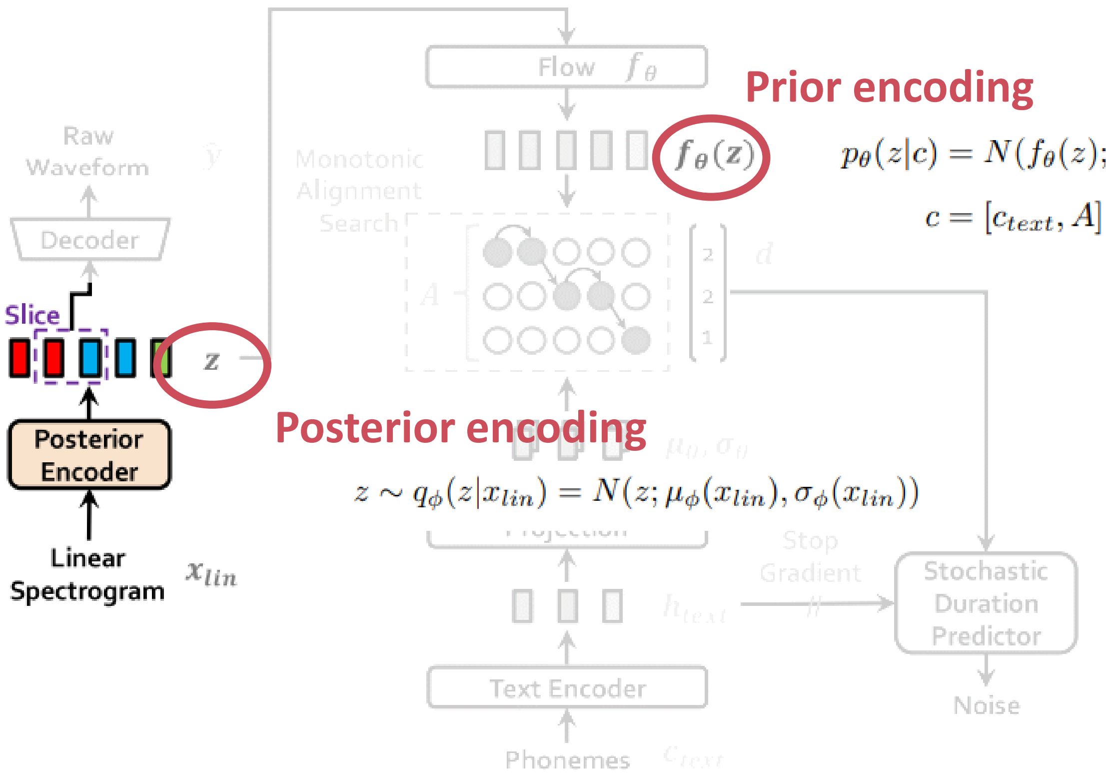
$$p_\theta(z|c) = N(f_\theta(z); \mu_\theta(c), \sigma_\theta(c)) \left| \det \frac{\partial f_\theta(z)}{\partial z} \right|,$$

$$c = [c_{text}, A]$$

$$L_{cvae} = L_{recon} + \underline{D_{KL}(q(z|y)||p(z|c))}$$

Fully end-to-end speech synthesis

VITS



Prior encoding

$$p_\theta(z|c) = N(f_\theta(z); \mu_\theta(c), \sigma_\theta(c)) \left| \det \frac{\partial f_\theta(z)}{\partial z} \right|,$$

$$c = [c_{text}, A]$$

Posterior encoding

$$z \sim q_\phi(z|x_{lin}) = N(z; \mu_\phi(x_{lin}), \sigma_\phi(x_{lin}))$$

$$L_{cvae} = L_{recon} + D_{KL}(q(z|y)||p(z|c))$$

Fully end-to-end speech synthesis

VITS

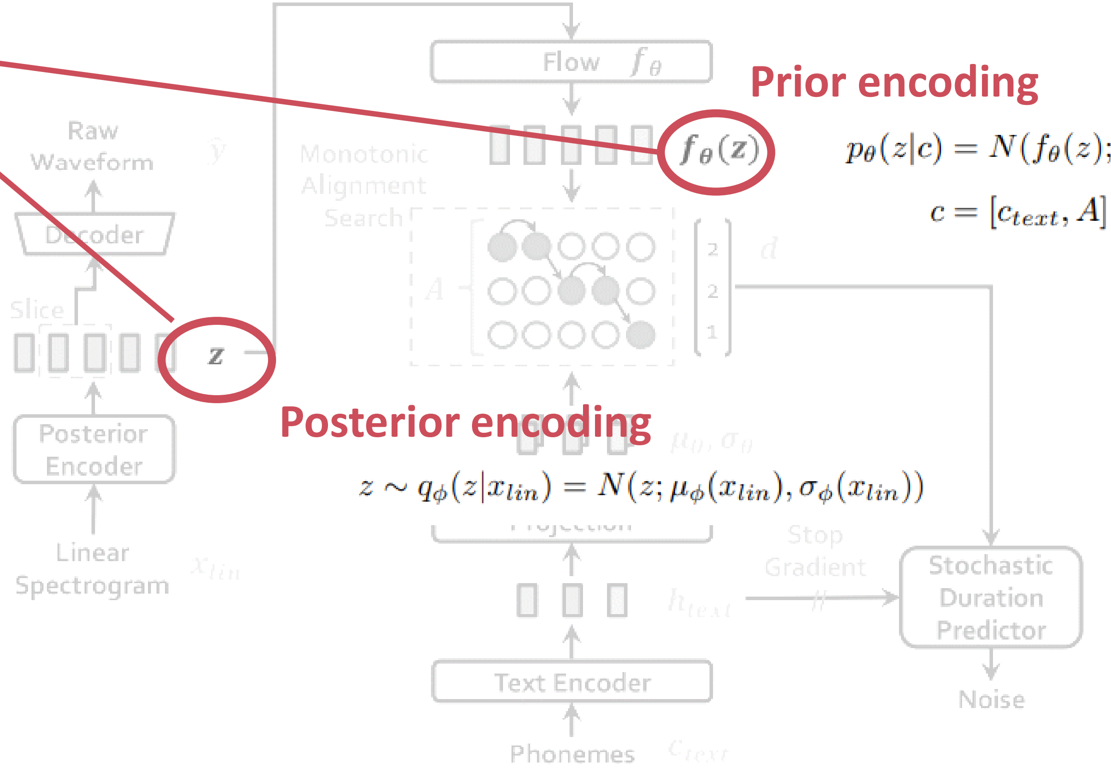
표현력 개선에 도움이 됨

Normalizing flow

Prior encoding

$$p_\theta(z|c) = N(f_\theta(z); \mu_\theta(c), \sigma_\theta(c)) \left| \det \frac{\partial f_\theta(z)}{\partial z} \right|,$$

$$c = [c_{text}, A]$$



Posterior encoding

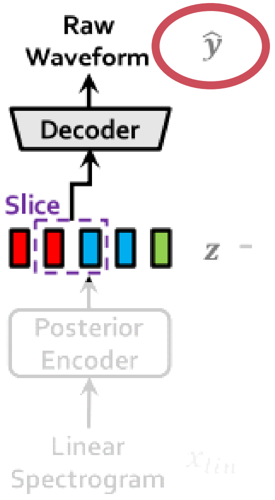
$$z \sim q_\phi(z|x_{lin}) = N(z; \mu_\phi(x_{lin}), \sigma_\phi(x_{lin}))$$

$$L_{cvae} = L_{recon} + \underline{D_{KL}(q(z|y)||p(z|c))}$$

Fully end-to-end speech synthesis

VITS

Synthesized speech



$$L_{recon} = \|x_{mel} - \hat{x}_{mel}\|_1$$

→ **Mel-spectrogram loss**

$$L_{adv}(D) = \mathbb{E}_{(y,z)}\Big[(D(y)-1)^2 + (D(G(z)))^2\Big],$$

$$L_{adv}(G) = \mathbb{E}_{z}\Big[(D(G(z))-1)^2\Big],$$

→ **Adversarial loss**

$$L_{fm}(G) = \mathbb{E}_{(y,z)}\Big[\sum_{l=1}^{T}\frac{1}{N_l}\|D^l(y) - D^l(G(z))\|_1\Big]$$

→ **Feature-matching loss**

$$L_{cvae} = \underline{L_{recon}} + D_{KL}(q(z|y)||p(z|c))$$

Fully end-to-end speech synthesis

VITS

Table 1. Comparison of evaluated MOS with 95% confidence intervals on the LJ Speech dataset.

Model	MOS (CI)
Ground Truth	4.46 ($\pm$0.06)
Tacotron 2 + HiFi-GAN	3.77 ($\pm$0.08)
Tacotron 2 + HiFi-GAN (Fine-tuned)	4.25 ($\pm$0.07)
Glow-TTS + HiFi-GAN	4.14 ($\pm$0.07)
Glow-TTS + HiFi-GAN (Fine-tuned)	4.32 ($\pm$0.07)
VITS (DDP)	4.39 ($\pm$0.06)
VITS	**4.43 ($\pm$0.06)**

Table 3. Comparison of evaluated MOS with 95% confidence intervals on the VCTK dataset.

Model	MOS (CI)
Ground Truth	4.38 ($\pm$0.07)
Tacotron 2 + HiFi-GAN	3.14 ($\pm$0.09)
Tacotron 2 + HiFi-GAN (Fine-tuned)	3.19 ($\pm$0.09)
Glow-TTS + HiFi-GAN	3.76 ($\pm$0.07)
Glow-TTS + HiFi-GAN (Fine-tuned)	3.82 ($\pm$0.07)
VITS	**4.38 ($\pm$0.06)**

1. Fine-tuning (w/ generated parameters) 도움이 됨

https://github.com/jaywalnut310/vits

Fully end-to-end speech synthesis

VITS

Table 1. Comparison of evaluated MOS with 95% confidence intervals on the LJ Speech dataset.

Model	MOS (CI)
Ground Truth	4.46 (±0.06)
Tacotron 2 + HiFi-GAN	3.77 (±0.08)
Tacotron 2 + HiFi-GAN (Fine-tuned)	4.25 (±0.07)
Glow-TTS + HiFi-GAN	4.14 (±0.07)
Glow-TTS + HiFi-GAN (Fine-tuned)	4.32 (±0.07)
VITS (DDP)	4.39 (±0.06)
VITS	**4.43 (±0.06)**

Table 3. Comparison of evaluated MOS with 95% confidence intervals on the VCTK dataset.

Model	MOS (CI)
Ground Truth	4.38 (±0.07)
Tacotron 2 + HiFi-GAN	3.14 (±0.09)
Tacotron 2 + HiFi-GAN (Fine-tuned)	3.19 (±0.09)
Glow-TTS + HiFi-GAN	3.76 (±0.07)
Glow-TTS + HiFi-GAN (Fine-tuned)	3.82 (±0.07)
VITS	**4.38 (±0.06)**

1. Fine-tuning (w/ generated parameters) 도움이 됨
2. **그래도 fully end-to-end 방법의 (VITS) 성능이 더 좋음**

https://github.com/jaywalnut310/vits

Speech synthesis and its applications

1. Speech analysis: Mel-spectrogram
2. Acoustic models: From text to acoustic parameters
3. Vocoder: From acoustic parameters to speech
4. Fully end-to-end speech synthesis
5. **Applications**

Whale Browser

Naver Dictionary

Navigation

Audio Book

Clova Speaker

Device

Ai Call

Care Call

+

'유인나' Voice
클로바 스피커 기본 목소리

‘오상진’ Voice
네이버 뉴스 본문 듣기 목소리


내비게이션 × Clova
내 차안의 인공지능 비서
980m
좌회전
600m
60
1km
980m
좌회전
600m
60
1 테라로사 커피공장
2 보헤미안박이추커피
3 순두부젤라또

변화의 시작! 꿈이 현실이 되는 해운대

클로바 케어콜이란?

인공지능 AI가 친구처럼 돌봄대상자와 안부를 묻고 답하는 서비스

AI와 어떤 대화를 할까요?

"간밤에 잠은 잘 주무셨어요?"

"아니오, 요즘 통 잠을 제대로 못자고 있네요"

"왜 그러세요? 무슨 걱정거리라도 있으세요?"

"그건 아닌데, 무릎이 아파서…"

"관절이 많이 안 좋으신가요?"

"부쩍 밤마다 많이 아프네요…"

"아이고 그러셨군요, 파스나 찜질팩이라도 해보세요.
병원에도 한번 가보시고요."

(어르신과 AI의 실제 대화 내용 중 발췌)

인공지능부터 로봇까지...네이버 실험실 거듭난 '1784'

이영아 기자 | 승인 2022.04.22 17:26

동영상에 보이스를 더하다

CLOVA Dubbing$^{\beta}$

자연스러운 클로바보이스로 동영상에
특별한 생동감을 더해주세요

Q / A

gregorio.song@gmail.com
eunwoo.song@navercorp.com