

Clova AI

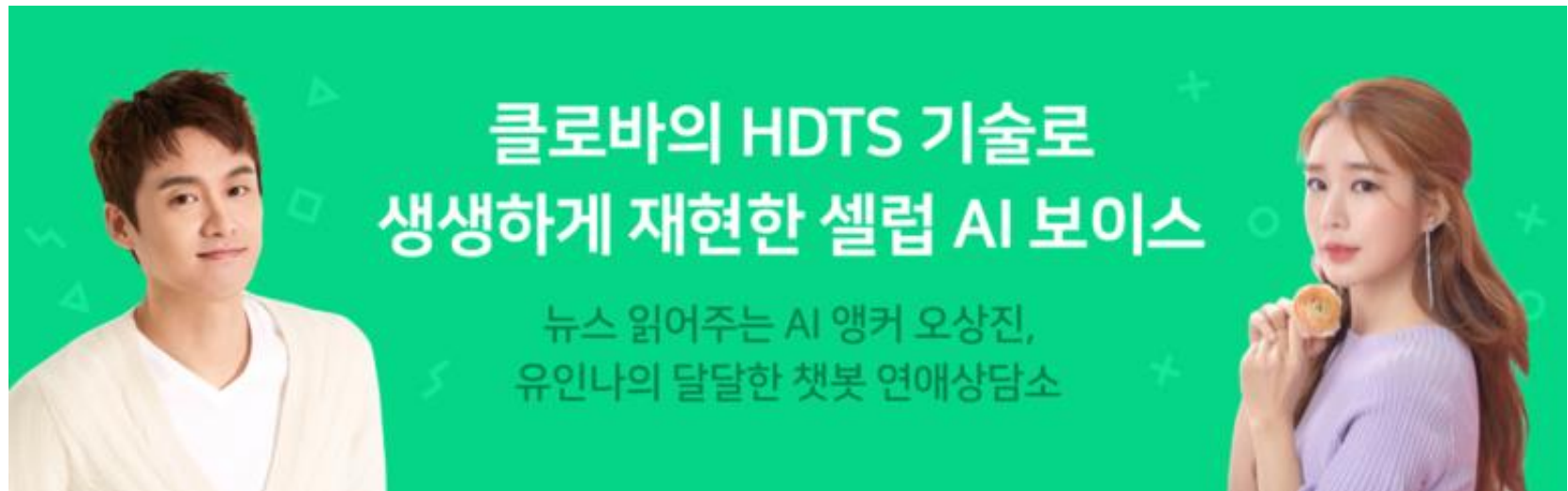
Text-to-Speech Technology

송은우 / HDTS 리더

발표내용

Text-to-speech (TTS)란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.
클로바의 음성합성 시스템은 자연스럽게 명확한 음성으로 좋은 평을 받고있으며
음성 대화 시스템, 파파고 자동 번역, 인공지능 스피커, 뉴스 읽어주기, 오디오 북 등
네이버의 다양한 서비스에 활용되고 있습니다.

이 강연에서는 **Unit-selection TTS** 시스템부터 **Statistical Parametric TTS** 시스템까지
클로바의 다양한 음성 합성 기술을 소개하고자 합니다.



Unit-selection TTS

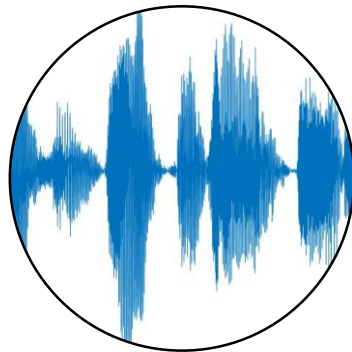
Unit-selection TTS

문맥에 맞는 유닛을 붙여서 음성을 만들자

Input
Text

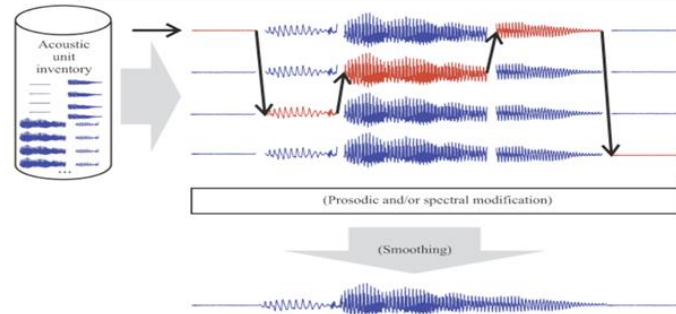
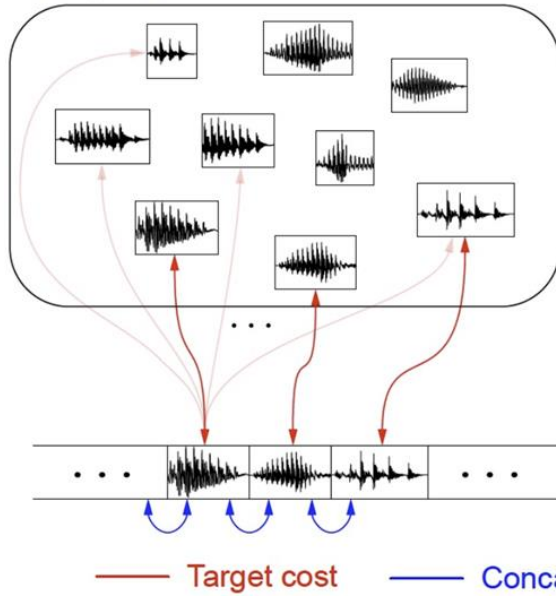
Unit-selection TTS

문맥에 맞는 유닛을 붙여서 음성을 만들자



Output Speech

All segments



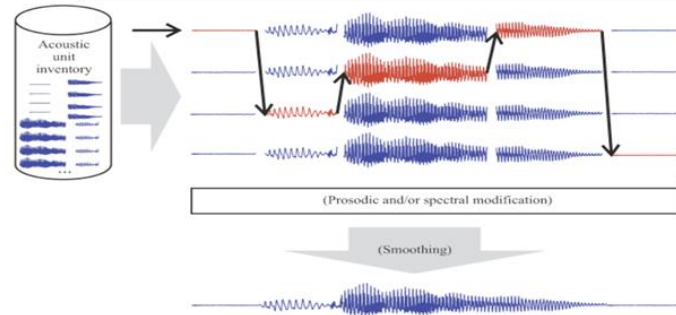
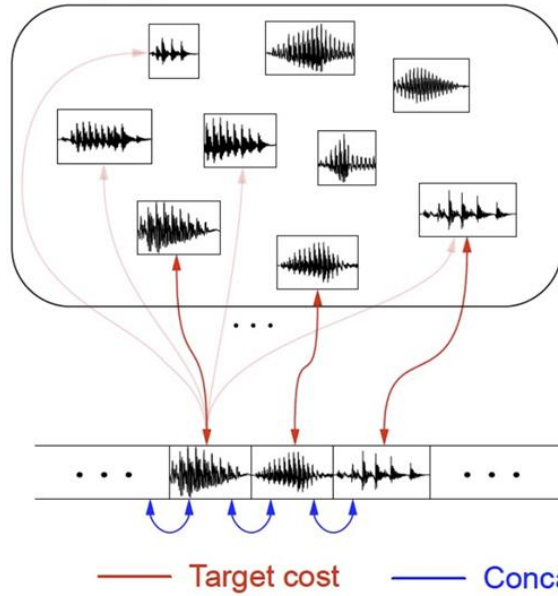
$$C(t_1^n, u_1^n) = \sum_{i=1}^n C^{target}(t_i, u_i) + \sum_{i=2}^n C^{join}(u_{i-1}, u_i)$$

$$\hat{u}_1^n = \operatorname{argmin}_{u_1, \dots, u_n} C(t_1^n, u_1^n)$$

그림 출처: T. Keiichi, and H. Zen. "Fundamentals and recent advances in HMM-based speech synthesis." *Tutorial of INTERSPEECH*, 2009.
 기술 출처: A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, 1996.

문맥에 맞는 유닛을 붙여서 음성을 만들자

All segments



$$C(t_1^n, u_1^n) = \sum_{i=1}^n C^{target}(t_i, u_i) + \sum_{i=2}^n C^{join}(u_{i-1}, u_i)$$

$$\hat{u}_1^n = \underset{u_1, \dots, u_n}{\operatorname{argmin}} C(t_1^n, u_1^n)$$

그림 출처: T. Keiichi, and H. Zen. "Fundamentals and recent advances in HMM-based speech synthesis." *Tutorial of INTERSPEECH*, 2009.

기술 출처: A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP* 1996.

지금까지도 모든 TTS 서비스에 사용되는 기술

Whale Browser



papago

Navigation



Clova Speaker



Naver Dictionary



Audio Book



네이버 TTS 서비스의 90%에 사용

Why ?

High-quality, Fast Generation

네이버 TTS 서비스의 90%에 사용

Why?

High-quality, Fast Generation

네이버 TTS 서비스의 90%에 사용

100hrs UTS 방식의 녹음 필요 시간

1.5yrs 약 1.5년여의 개발 기간

최고 품질의 음성 합성 기술력

100hrs UTS 방식의 녹음 필요 시간

1.5yrs 약 1.5년여의 개발 기간

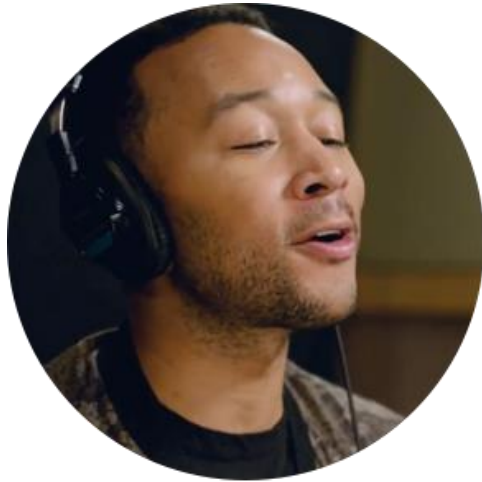
다음 단계로 나아가기 위해 필요한 것은?



해투4 서유리 “빅스비 내 목소리,
1년 녹음→돈 많이 받았다”
(2019.05.02, 뉴스엔)

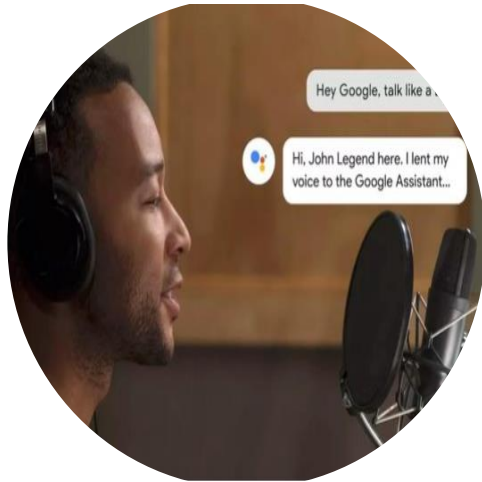


서유리는 “1년 동안 하루 4시간씩 주 5회 녹음을 했다. 돈을 많이 받았다”고 말해 출연자들의 관심을 한몸에 받았다. 서유리는 “비밀 유지 계약을 써서 몇 단어를 녹음했는지는 말을 못 한다. 한마디만 하자면 요구하는 부분이 상당히 많았다”고 덧붙였다.



구글 어시스턴트에
존 레전드 목소리 적용
(2018.05.19, 아주경제)

구글의 음성 인식 AI 비서 '구글 어시스턴트'에 미국의 R&B, 소울 싱어송라이터 **존 레전드**의 목소리가 적용된다...구글의 자회사 **딥마인드**가 개발한 **웨이브넷(WaveNet)**을 이용해 자연스러운 목소리가 구현됐다.



존 레전드 외 5명 구글 음성인식 비서,
유명 연예인 목소리로 서비스
(2019.04.04, 중앙일보)

3일(현지시간)부터 구글의 인공지능(AI) 음성인식 비서
'구글 어시스턴트'가 유명 연예인의 목소리로 사용자에게
필요한 정보를 알려주는 서비스를 시작했다.

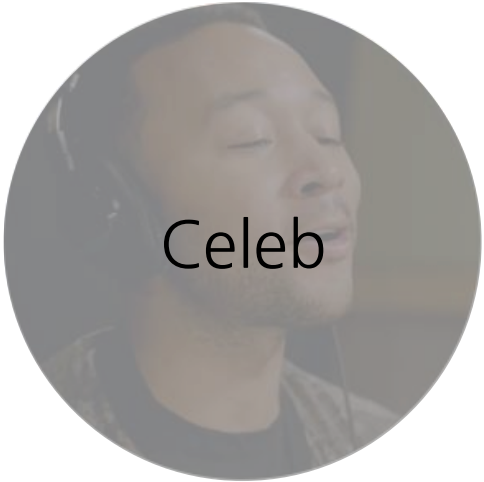
...

다만 초기 단계에는 **극히 제한된 질문**에만 유명인이
대답한다.



전문 성우

100시간의 녹음



Celeb

100시간의 녹음?

New Voice,
New Technology

HDTs

(High-quality DNN Text-to-Speech)

음성 접합

DB 확보

Concatenative

자연스러움

Unit-Selection

고된 녹음

대용량

실시간 합성

UTS

Generative

오랜 생성 시간

Adaptation

WaveNet

로봇 같은 소리

딤러닝

저비용

음성 생성

DTS

HDTTS

딥러닝 기반 음성합성 시스템

통계 모델을 활용하여 음성 신호 생성

Input
Text

딥러닝 기반 음성합성 시스템

통계 모델을 활용하여 음성 신호 생성



Output Speech



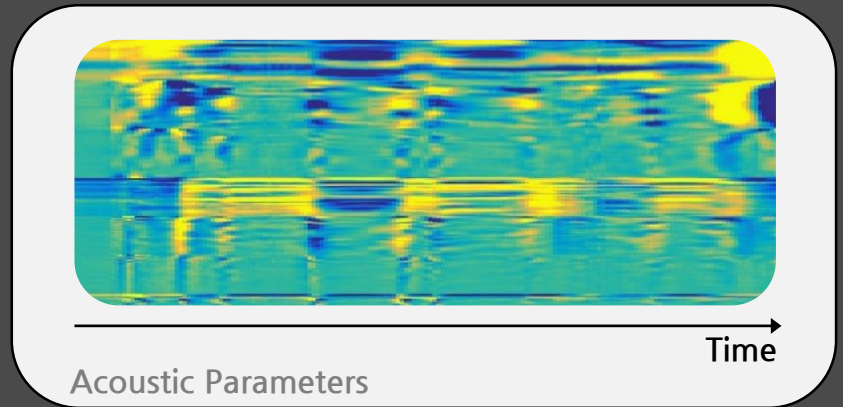
Acoustic Model + Vocoding Model

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

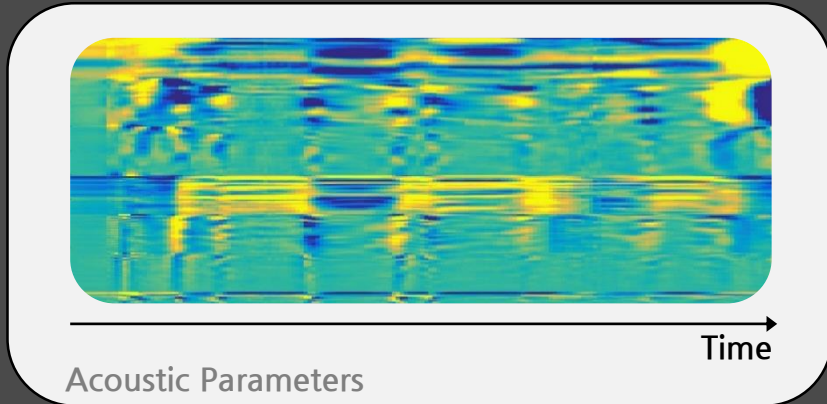


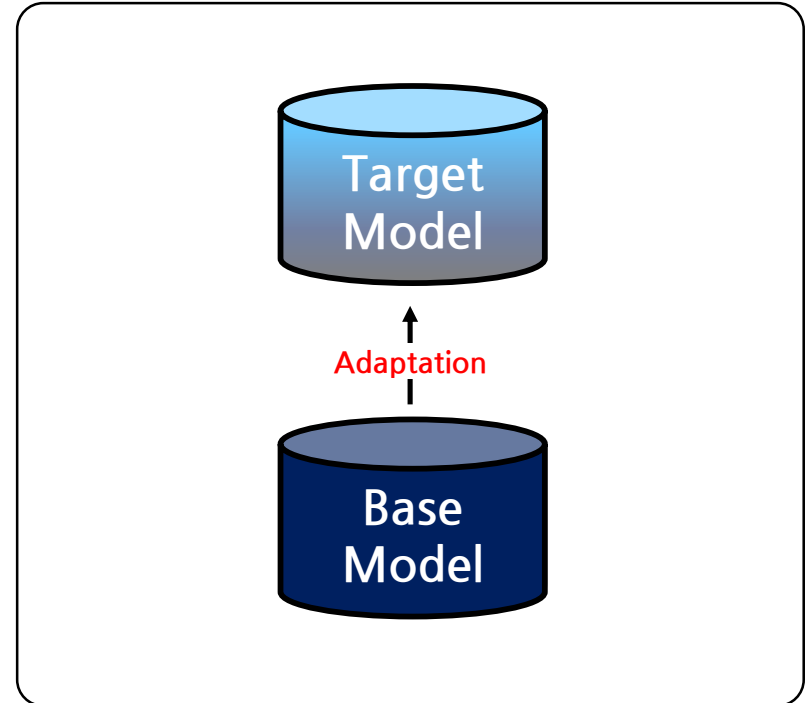
샐리야, 안녕?

Input Text



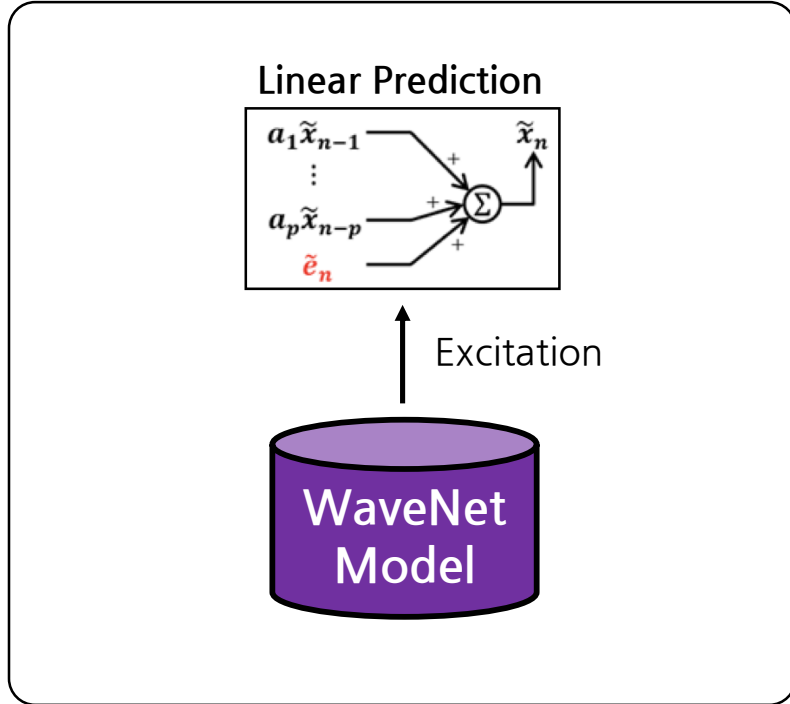
Acoustic Parameter 에서 음성 신호를 추정





1. Speaker Adaptation

100hrs → 4hrs (녹음 데이터 4% 이하로 감소)



2. LP-WaveNet Synthesis

합성음 품질 2배 향상 (MOS 2.3 → 4.5)



3. UTS + DTS 구조 결합

CPU 환경 - 1초 음성을 만들때 0.01초 소요

Speaker Adaptation

DB 녹음 4시간

LP-WaveNet 보코더

MOS 4.5

UTS + DTS 구조 결합

CPU + 0.01초

개발 비용 95% 감소

개발 기간 90% 감소



+



‘유인나’ Voice

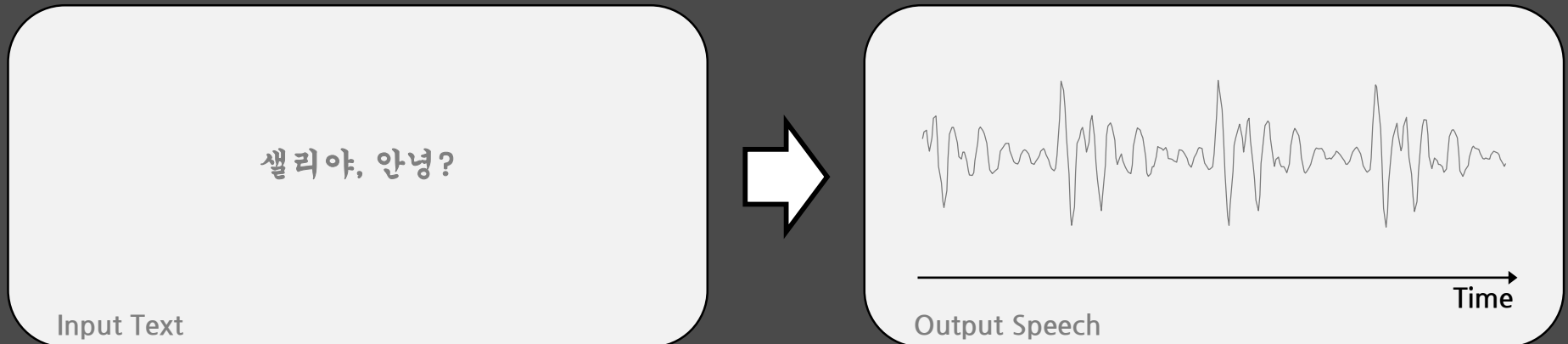
클로바 스피커 기본 목소리 적용 (2018.11)

Clova HDTS

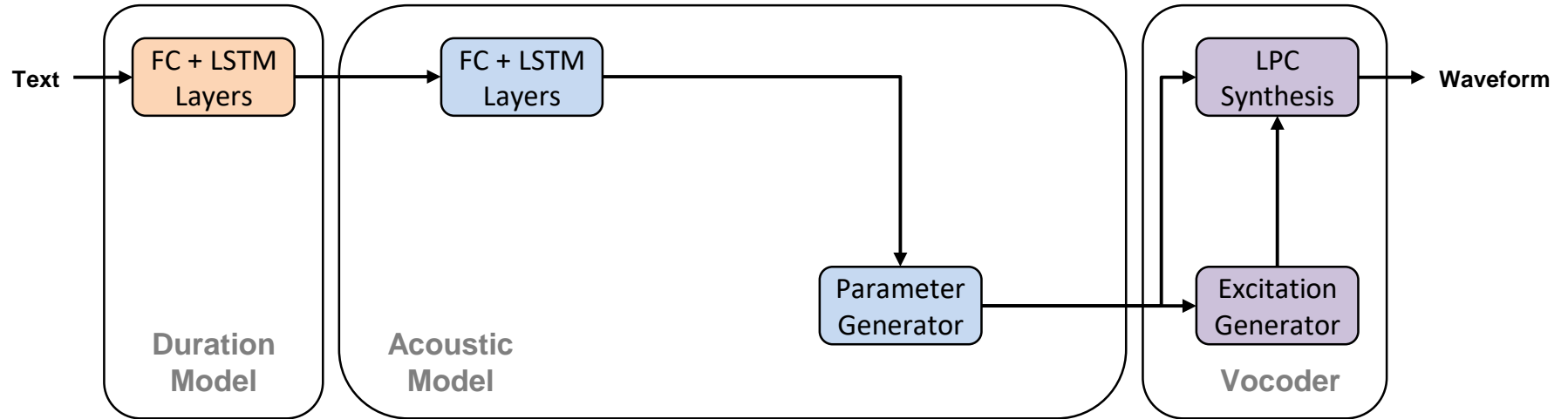
톤의 높낮이, 음색, 어조, 강세 등
텍스트에서 **Acoustic Parameter** 를 추정



Acoustic Parameter 에서 음성 신호를 추정



Clova HDTs



Duration model

음소열(phoneme sequence)의 시간축 길이(duration)을 추정하는 역할

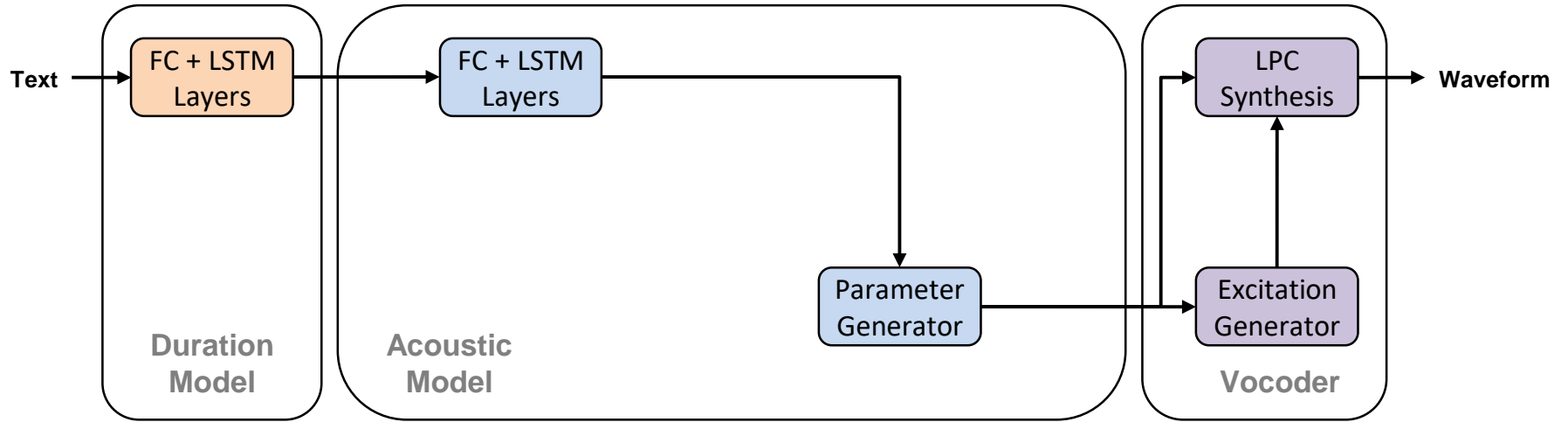
Acoustic model

텍스트 파라미터로부터 음성 파라미터(acoustic parameter)를 추정하는 역할

Vocoder

추정된 음성 파라미터로부터 음성신호(waveform)를 복원하는 역할

Clova HDTs



Duration model

음소열(phoneme sequence)의 시간축 길이(duration)를 추정하는 역할

Acoustic model

텍스트 파라미터로부터 음성 파라미터(acoustic parameter)를 추정하는 역할

Vocoder

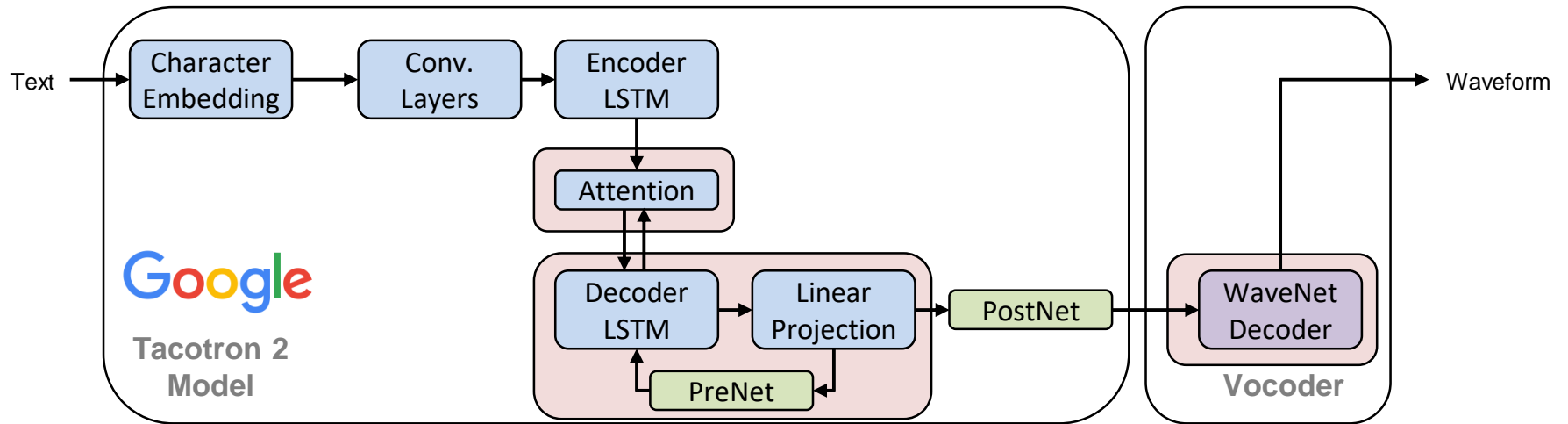
추정된 음성 파라미터로부터 음성신호(waveform)를 복원하는 역할

모델 학습을 위해서 각 음소(phoneme)의 시간축 길이(duration)를 알고 있어야 함
추정 과정에서 음성 파라미터(acoustic parameter)가 over-smoothing 됨
고도화된 신호처리(signal processing) 기술이 필요함



비싸요
소리가 나빠요
만들기 어려워요

Clova HDTs



Seq2seq model with attention
 Autoregressive acoustic model
 WaveNet vocoder

음소의 시간축 길이 없어도됨
 Over-smoothing 완화됨
 CNN 모델이 음성 생성해줌

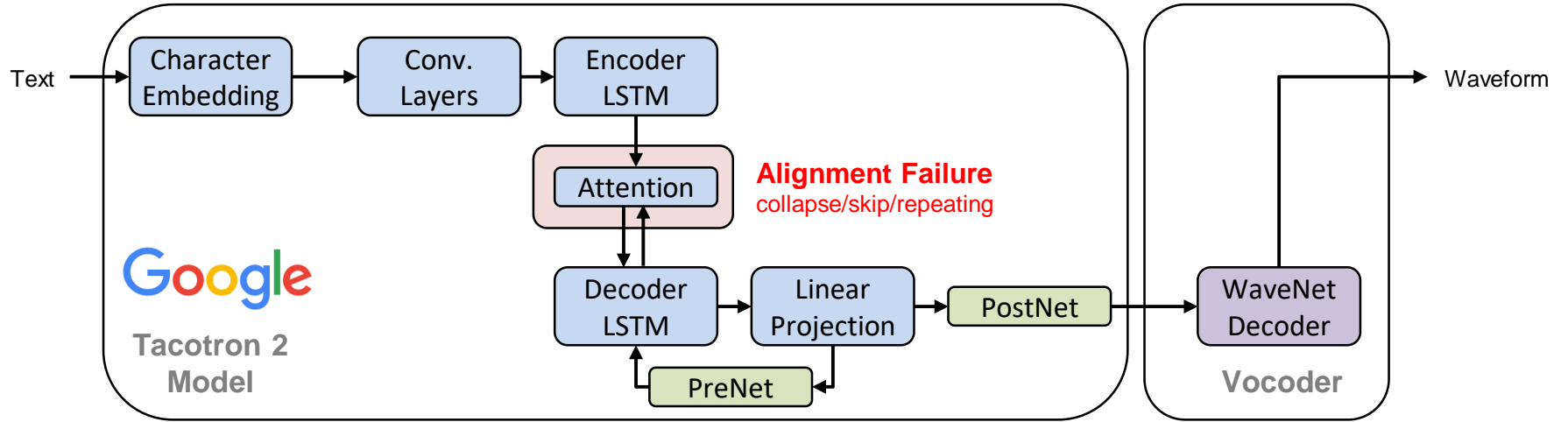


고품질의 음성 합성 엔진을 쉽게 만들 수 있게 됨



[이전 모델] [Tacotron 2]

Clova HDTS



Seq2seq model with attention
 Autoregressive acoustic model
 WaveNet vocoder

음소의 시간축 길이 없어도됨
 Over-smoothing 완화됨
 CNN 모델이 음성 생성해줌

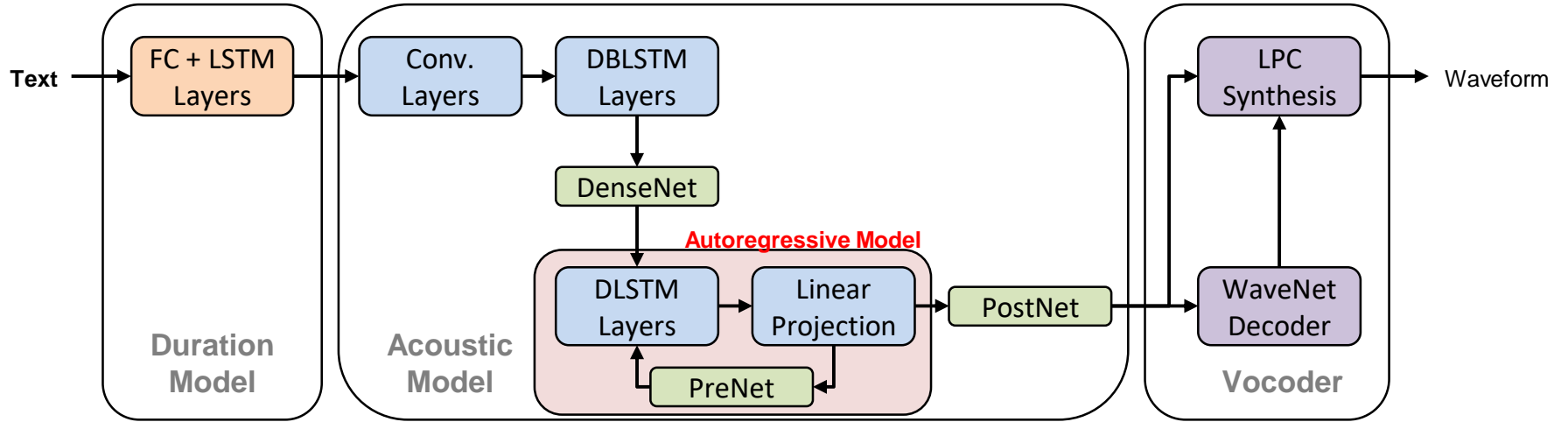


Attention 모델이 제대로 동작하지 않을 때, 합성음 품질에 Critical 영향을 줌



[이전 모델] [Tacotron 2]

Clova HDTS



[HDTS 음성 합성 모델]

External duration model

Autoregressive acoustic model

LP-WaveNet vocoder

음소의 시간축 길이는 추정하고

Tacotron decoder 구조는 유지하면서

WaveNet 모델의 성능까지 올린다.



고품질의 음성을 안정적으로 생성할 수 있게됨



[HDTS]



[이전 모델]



[Tacotron 2]

‘오상진’ Voice

네이버 뉴스 본문 듣기 적용 (2020.05)



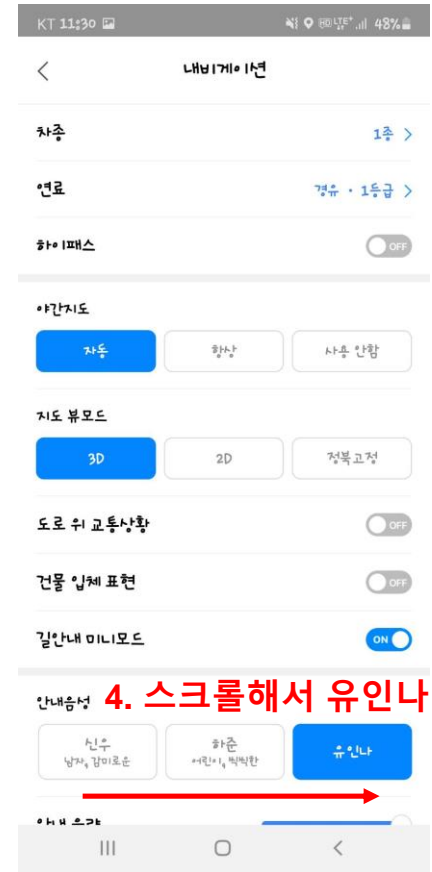
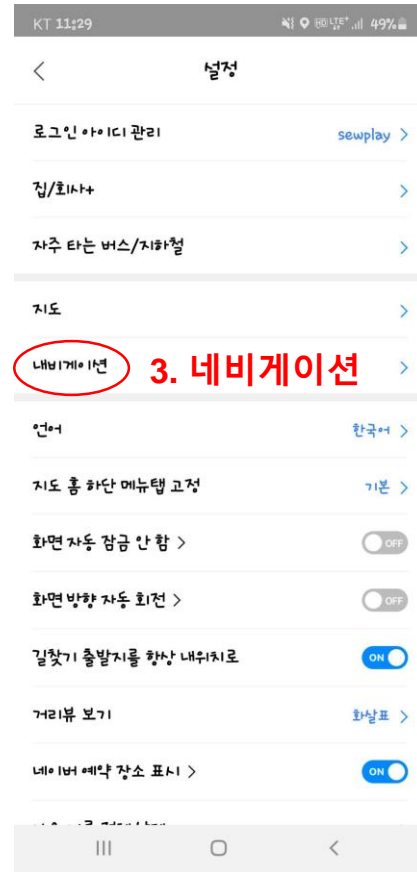
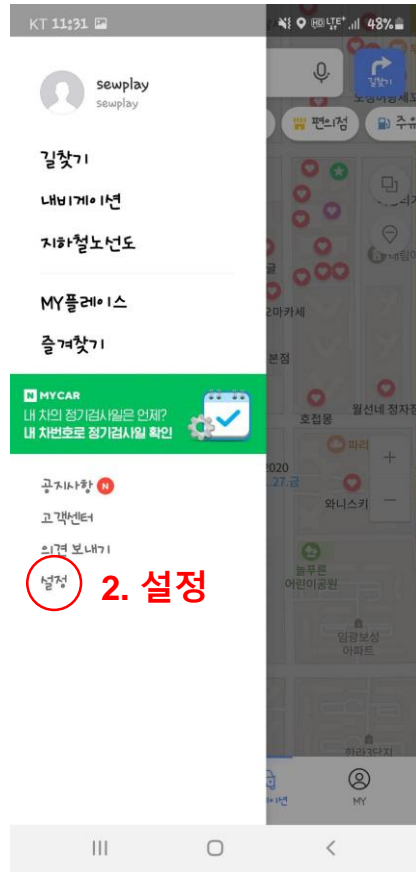
클로바의 HDTS 기술로
생생하게 재현한 셀럽 AI 보이스

뉴스 읽어주는 AI 앵커 오상진,
유인나의 달달한 챗봇 연애상담소



‘유인나’ Voice

네이버 지도 내비게이션 안내 (2020. 09)



‘유인나’ Voice
네이버 지도 내비게이션 안내 (2020. 09)

쉬는 시간

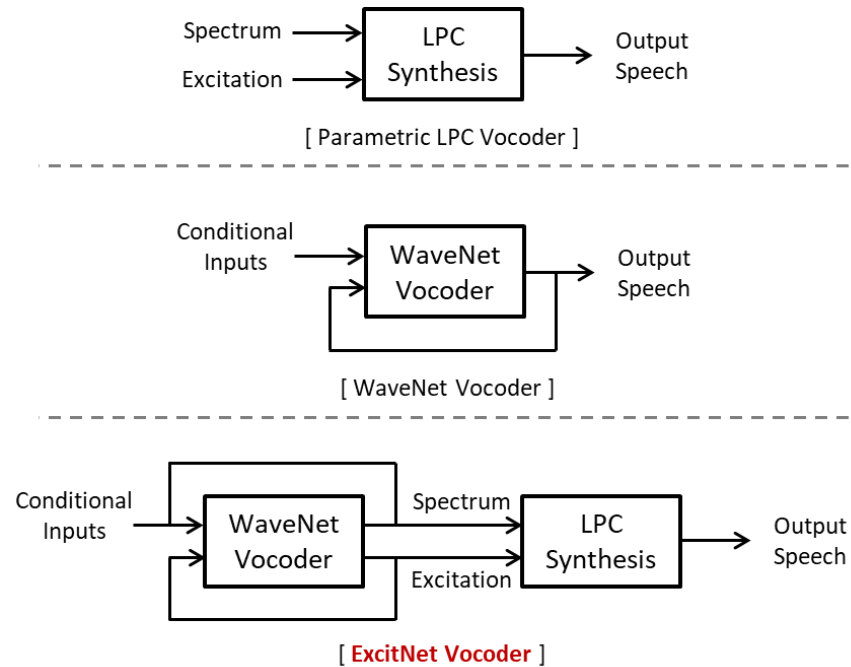


10분 뒤에 만나요

HDTTS Research



Neural LPC Vocoder (1/4)



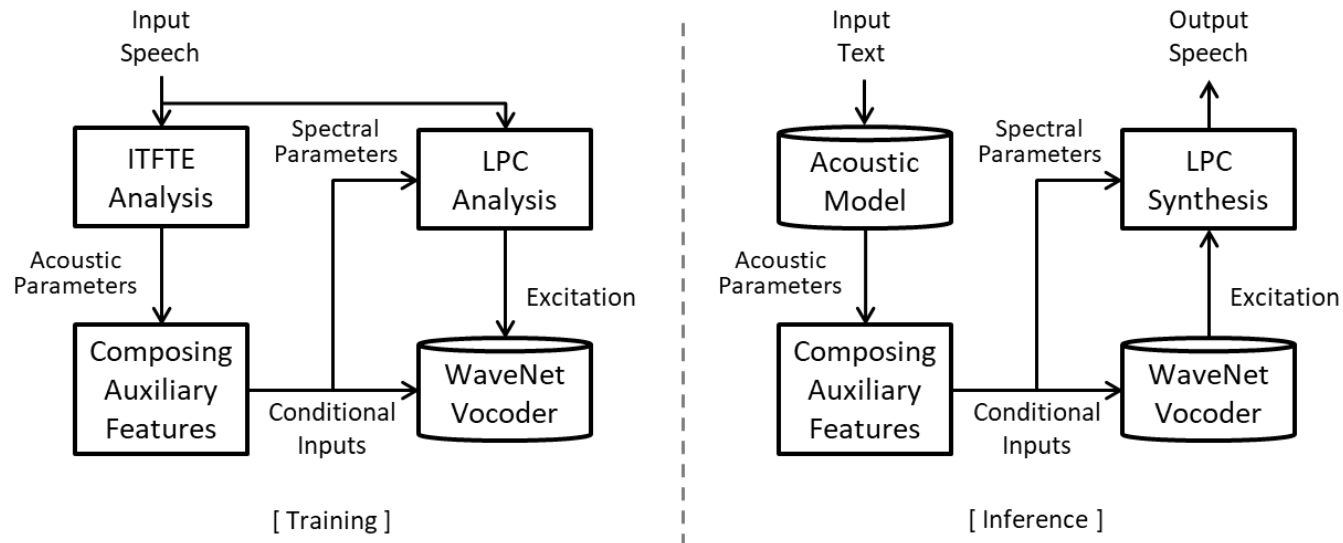
E. Song, K. Byun, H.-G. Kang, "ExcitNet vocoder: A neural excitation model for parametric speech synthesis systems," in *Proc. EUSIPCO*, 2019, pp. 1-5.

WaveNet Vocoder
Excitation generator

+

Linear Prediction Filter
Speech synthesis

Neural LPC Vocoder (2/4): ExcitNet



E. Song, K. Byun, H.-G. Kang, "ExcitNet vocoder: A neural excitation model for parametric speech synthesis systems," in *Proc. EUSIPCO*, 2019, pp. 1-5.

Korean Male Speaker

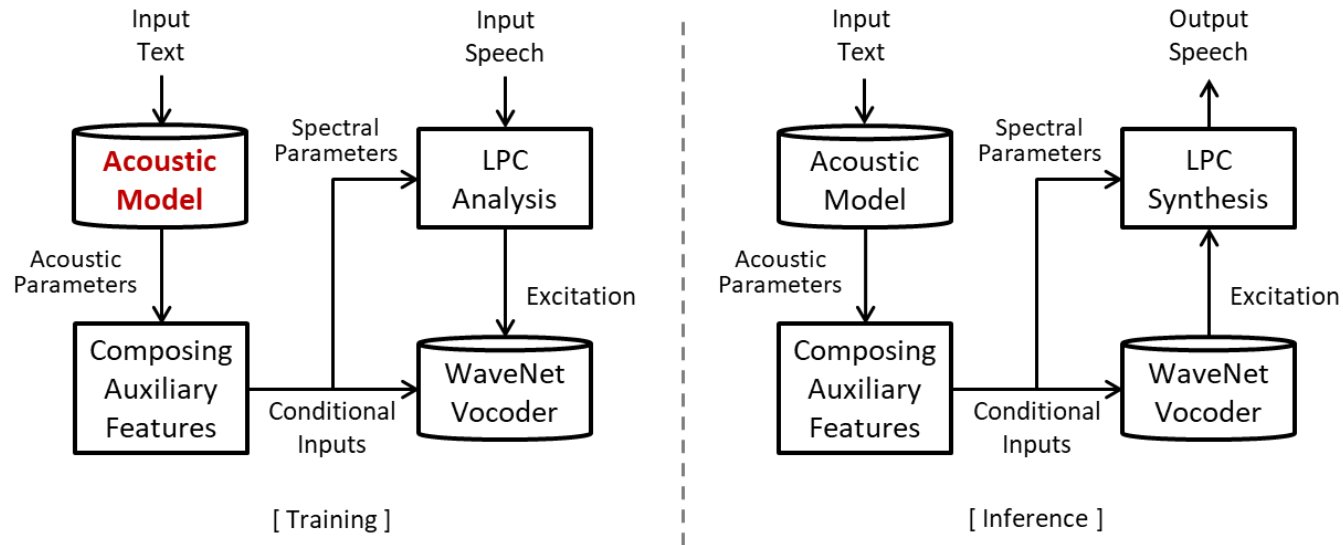
7 hours' recording

MOS Test Results

Recording: 4.58 

TTS: 3.99 

Neural LPC Vocoder (3/4): Modeling-by-Generation Training



E. Song, M.-J. Hwang, R. Yamamoto, O. Kwon, J. Kim, "Neural text-to-speech with a modeling-by-generation excitation vocoder," in *Proc. INTERSPEECH*, 2020, pp.3570-3574.

Korean female Speaker

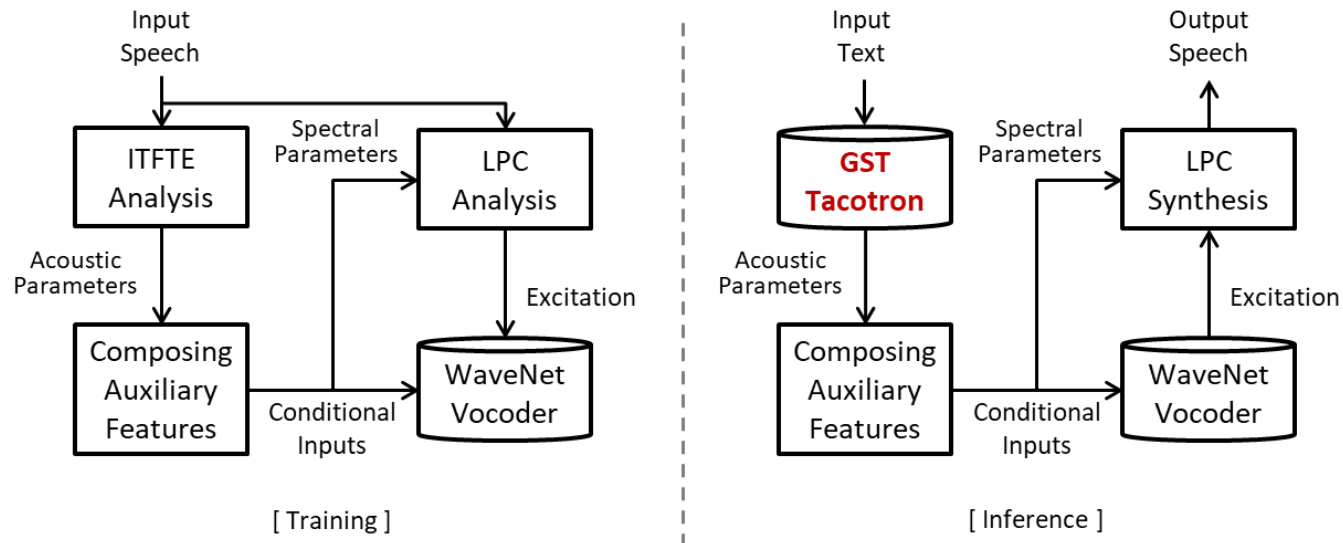
8 hours' recording

MOS Test Results

Recording: 4.66 

TTS: 4.57 

Neural LPC Vocoder (4/4): Emotional TTS



O. Kwon, E. Song, J.-M. Kim, H.-G. Kang, "Effective parameter estimation methods for an ExcitNet model in generative text-to-speech systems," *arXiv preprint arXiv:1905.08486*, 2019.

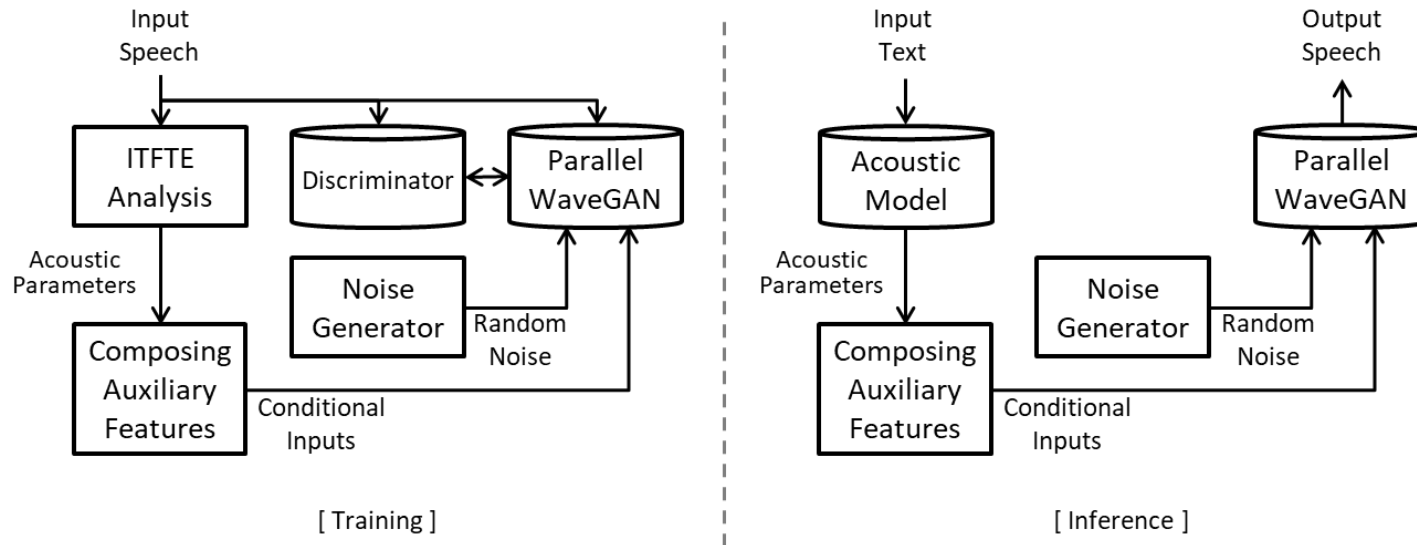
Korean female Speaker
4 hours' recording / emotion

MOS Test Results

Happy (TTS): 4.67 

Sad (TTS): 4.43 

Parallel WaveGAN Vocoder (1/3)



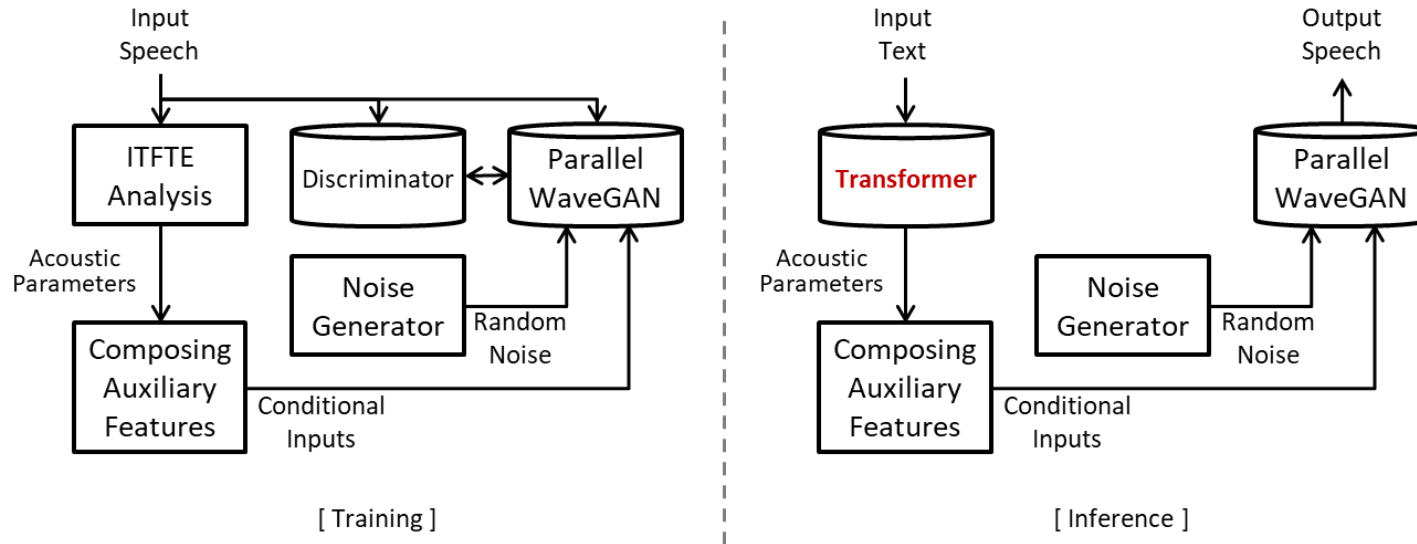
R. Yamamoto, E. Song, J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, 2020, pp. 6194-6198.

**Non-Autoregressive
WaveNet**
Fast generation

+

Adversarial Training
Effective Training without
teacher-student Framework

Parallel WaveGAN Vocoder (2/3): End-to-End TTS




R. Yamamoto, E. Song, J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, 2020, pp. 6194-6198.

Japanese Female Speaker

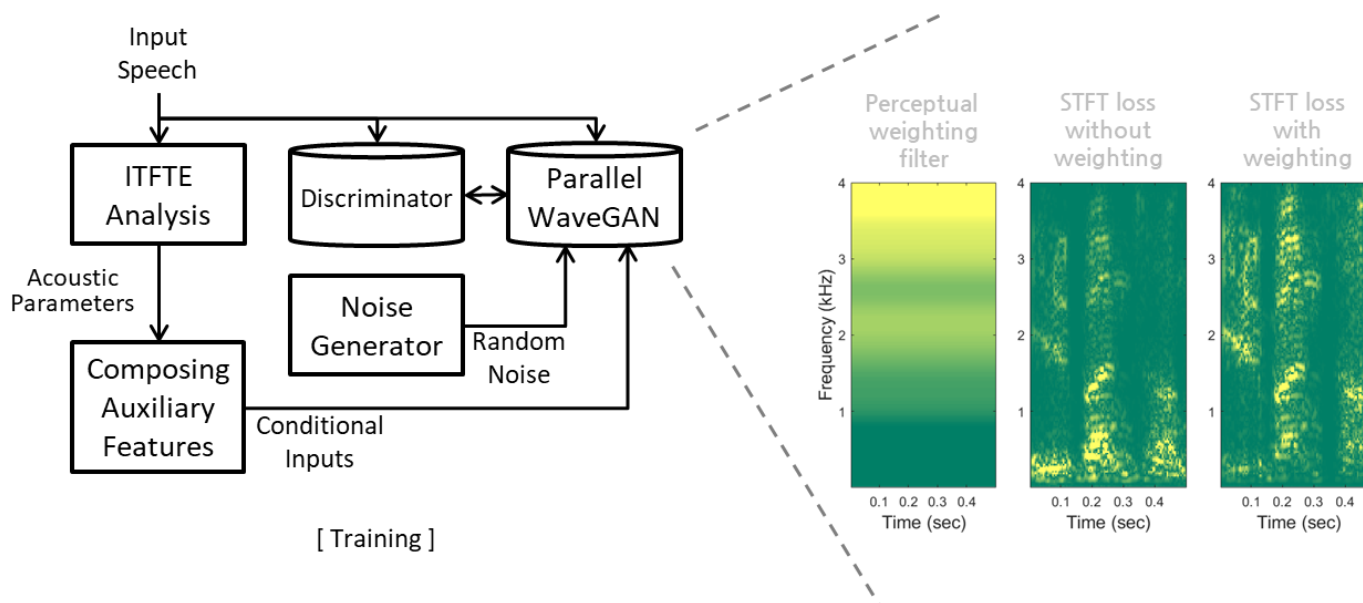
23 hours' recording

MOS Test Results

Recording: 4.46 

TTS: 4.16 

Parallel WaveGAN Vocoder (3/3): Perceptually Weighted Loss



E. Song, R. Yamamoto, M.-J. Hwang, O. Kwon, J. Kim, "Improved Parallel WaveGAN with perceptually weighted spectrogram loss," *accepted to Proc. SLT*, 2021.

Korean Male Speaker

7 hours' recording

MOS Test Results

Recording: 4.56 

TTS: 4.21 

Q / A



기술 제휴 문의
Kjm.kim@navercorp.com

채용 문의
eunwoo.song@navercorp.com

