DEVIEW 2021

클로바 보이스의 주인공을 찾습니다!

(누구나 만드는 내 목소리 합성기 III)

임선미

NAVER / NES / TL

CONTENTS

1.음성 합성이란?

음성 합성이란?

NEWS

NAVER CLOUD PLATFORM

동영상에 보이스를 더하다
CLOVA Dubbing β

Dict

papago

기술과 함께 하는
오디오클립

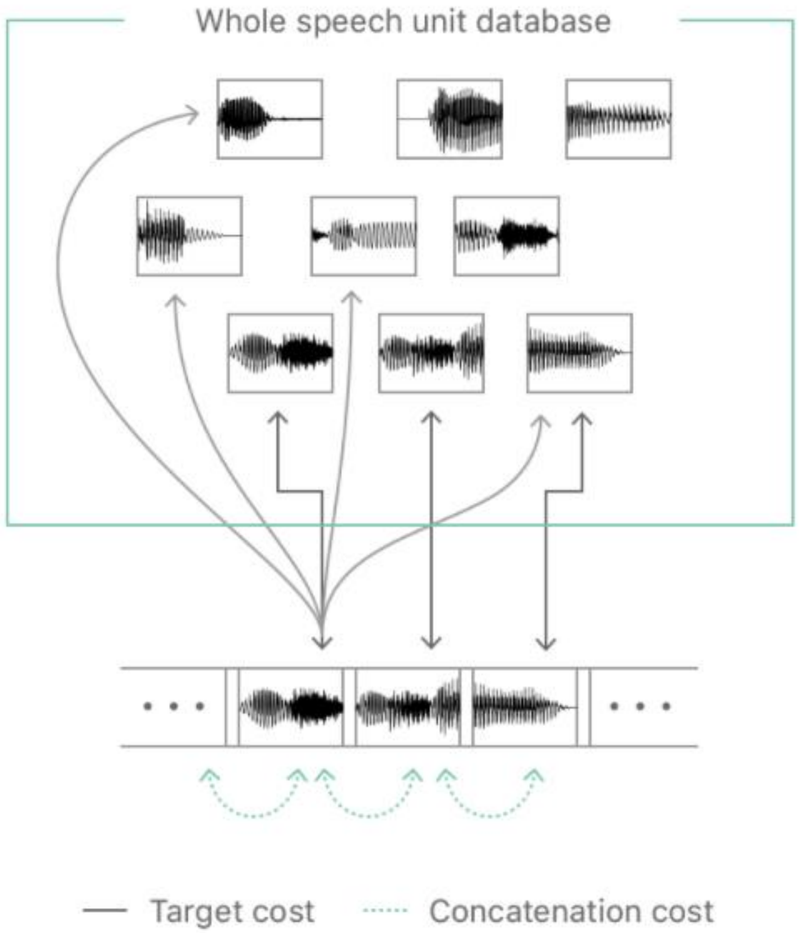
음성 합성이란?

DEVIEW 2021

Speech Synthesis

인공적으로 사람의 음성을 만드는 것

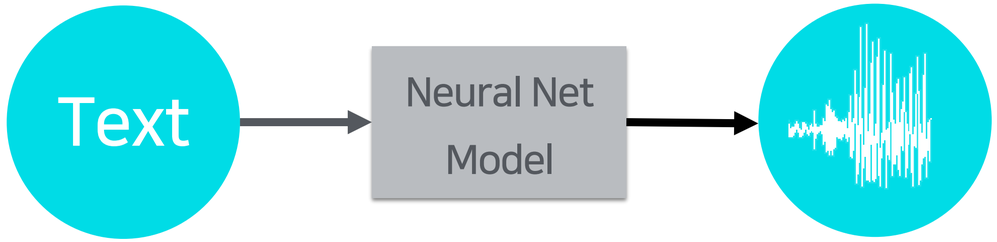
Concatenate 합성기



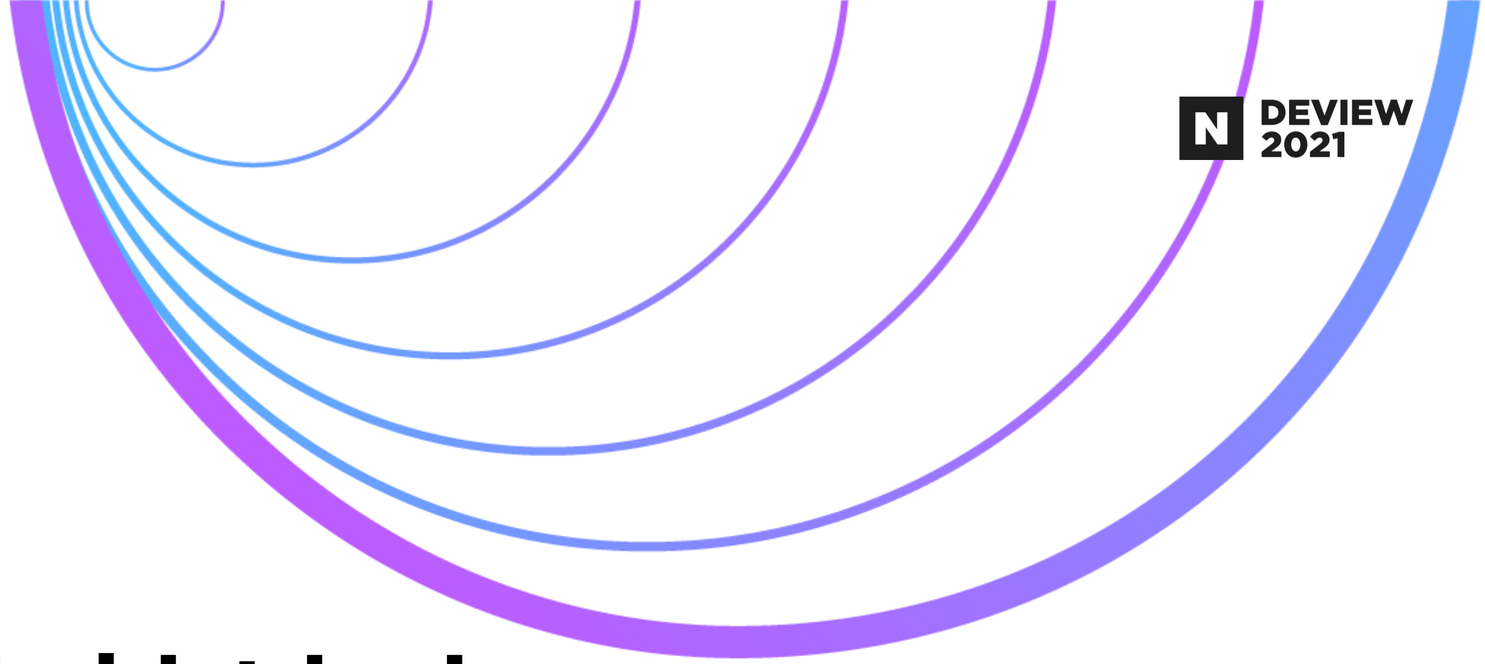
- 한 화자의 대용량 DB 필요
- 전사 과정 필요
- 추정이 잘 되면 원음 그대로의 고품질의 소리 생성 가능

https://machinelearning.apple.com/research/siri-voices

End-to-end 합성기



- 한 화자의 적은 DB
- 전사 과정 필요 X
- Neural Vocoder로 고품질 합성음 생성 가능

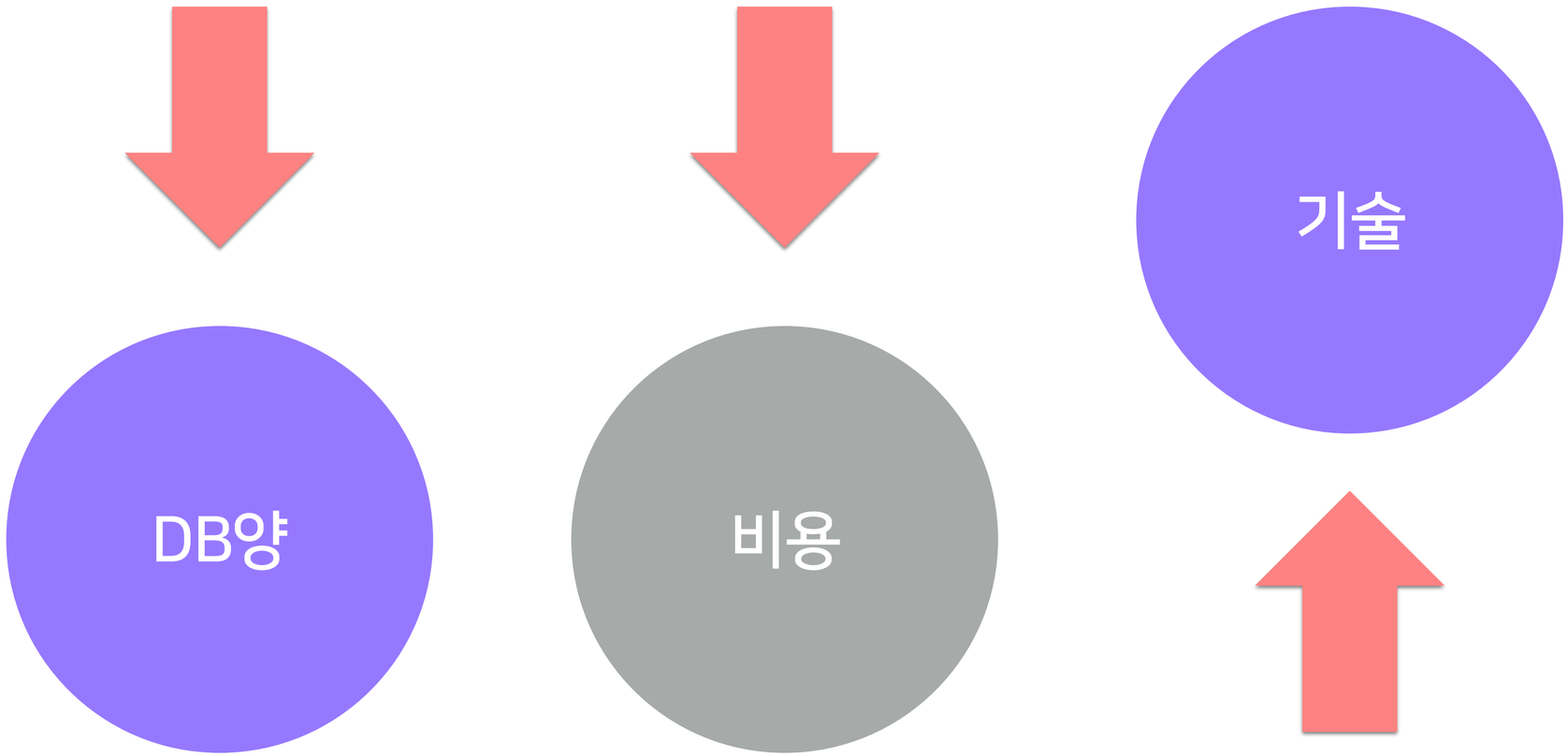


2. 개인화 합성기 History

개인화 합성기를 만들려면

해결해야 할 것들!!

개인화 합성기를 만들려면



개인화 합성기의 첫걸음

DB양

적은 데이터 사용

비용

전사 작업 X

스튜디오 녹음 X

기술

Neural acoustic model

Parametric vocoder 사용

개인화 합성기 서비스의 가능성 확인

DB양

적은 데이터 사용

비용

전사 작업 X

스튜디오 녹음 X

기술

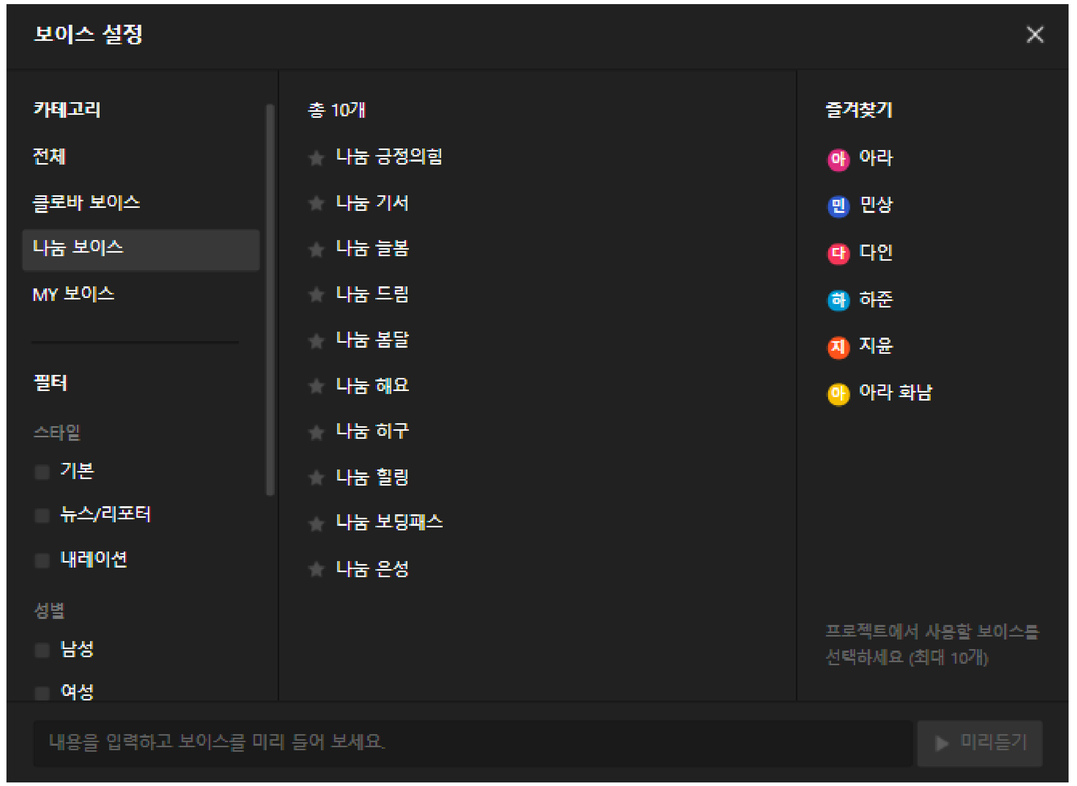
Neural acoustic model

Neural vocoder *

누구나 만드는 내 목소리 합성기 II (커스텀 보이스 파이프라인)

DEVIEW 2021

클로바더빙



2021년 현재,

개인화 합성기 서비스 준비

DB양

적은 데이터 사용

비용

전사 작업 X

스튜디오 녹음 X → 휴대폰 녹음 + 저품질 음성 탐지 *

기술

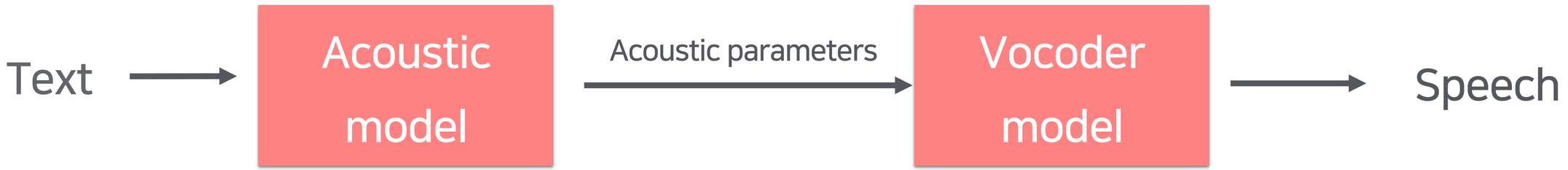
Neural acoustic model 고도화 *

Neural vocoder

데이터 전처리부터 모델 학습, 모델 선정까지 전 과정 자동화 시스템 구축 *

3. NES 개인화 합성기 구조

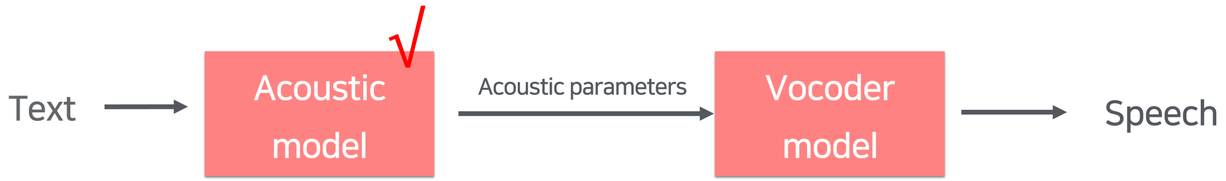
NES : Natural End-to-end Speech Synthesis System



Acoustic Model

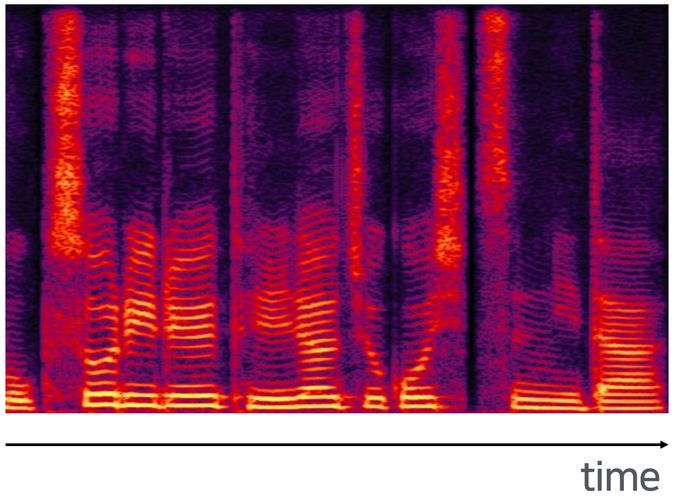
Text로부터 Acoustic parameters를 추정

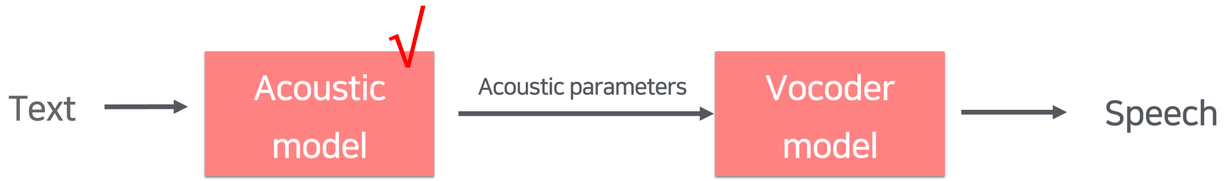
Text	Acoustic parameters
grapheme (자소) phoneme (음소)	mel spectrogram specific vocoder parameter



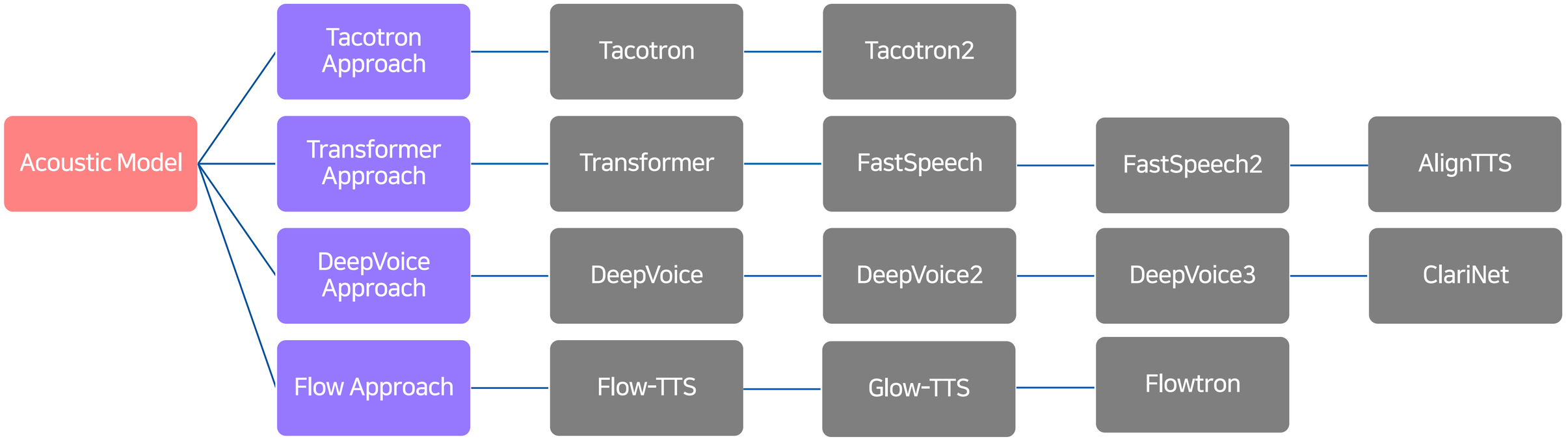
Acoustic Model

목소리를 들려주고
싶습니다.



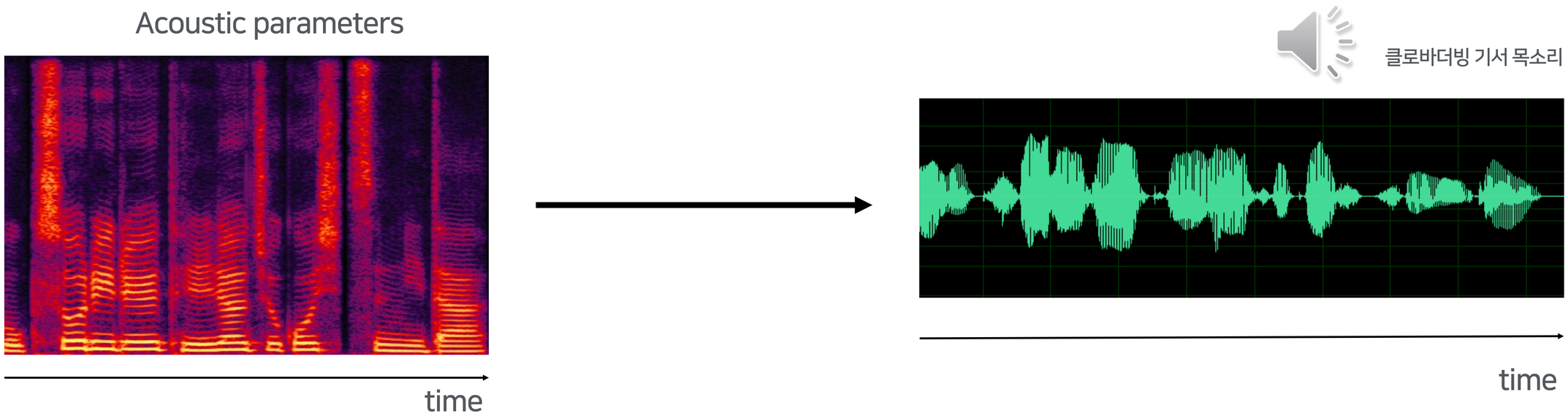


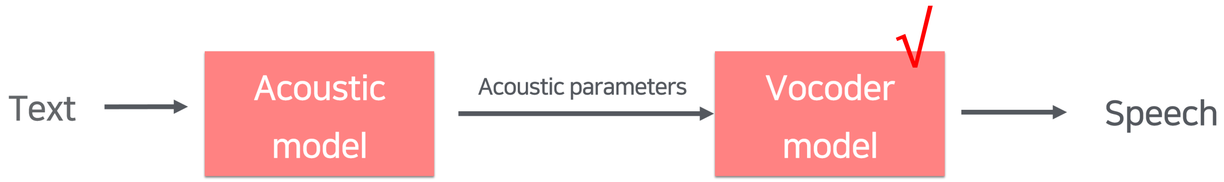
Acoustic Model



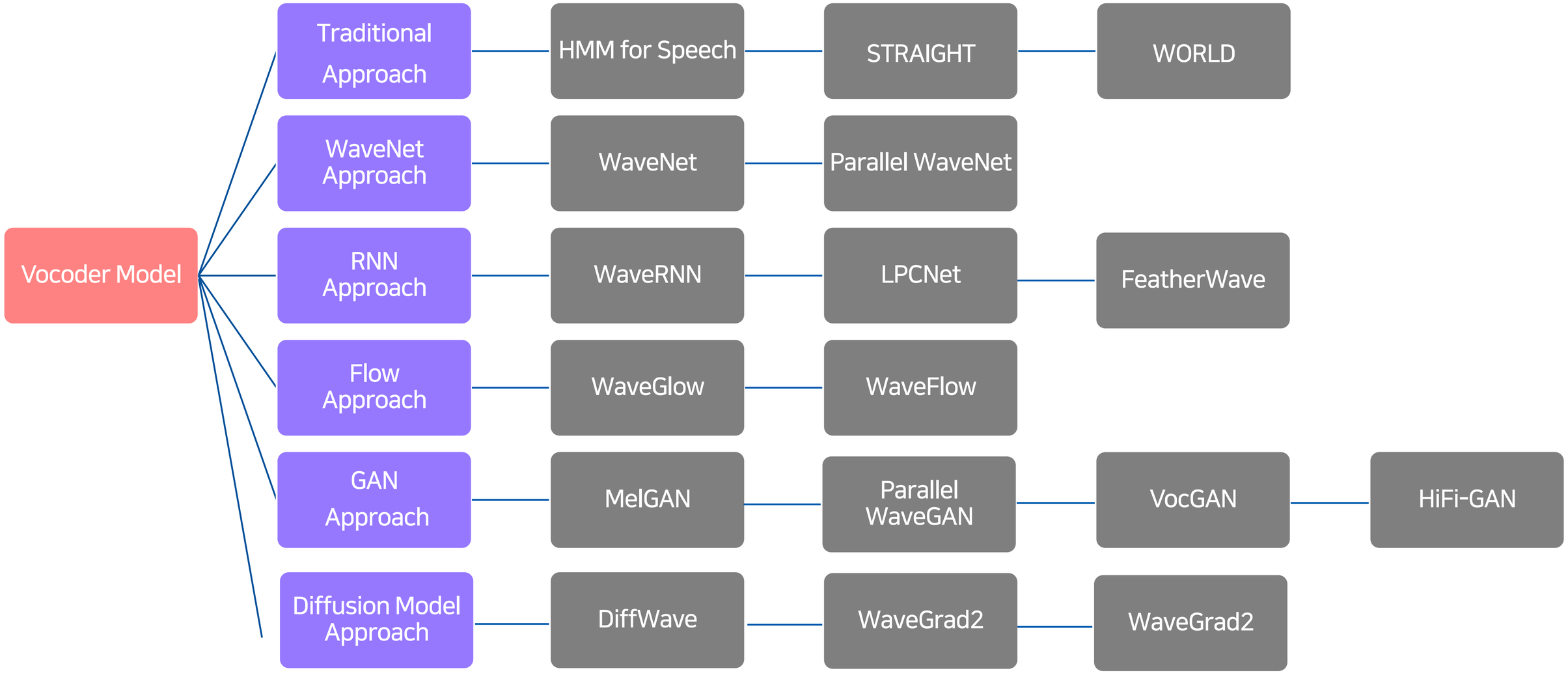
Vocoder Model

Acoustic parameters에서 음성 신호를 생성

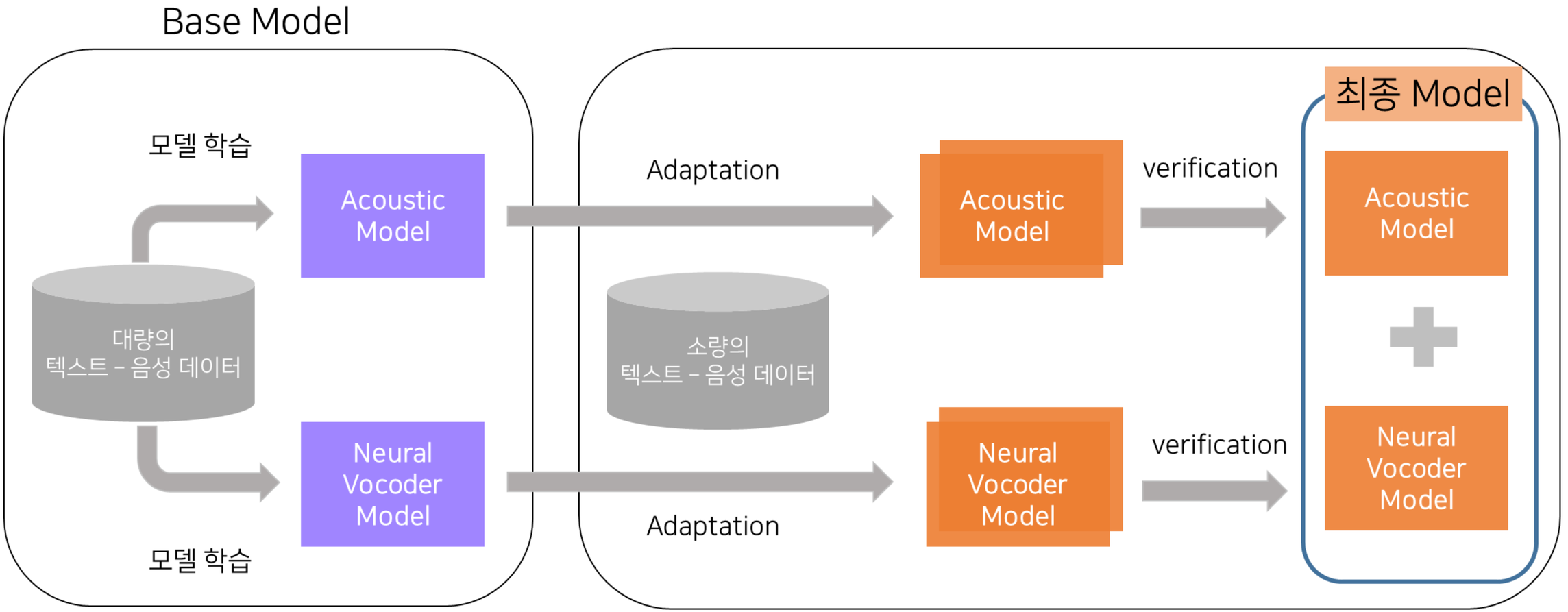




Vocoder Model



전체 학습 과정


Base Model
모델 학습
Acoustic Model
대량의 텍스트 – 음성 데이터
Neural Vocoder Model
모델 학습
Adaptation
소량의 텍스트 – 음성 데이터
Adaptation
Acoustic Model
Neural Vocoder Model
verification
verification
최종 Model
Acoustic Model
Neural Vocoder Model

4. 개인화 합성기 끝판왕
- Voice Maker 서비스

Voice Maker 서비스 탄생 배경

나눔 보이스 목소리가 자연스럽고 사람 답고(?) 좋더라구요.

고객의 목소리

나눔AI보이스 모집 때 일반인인데도 좋은 목소리들이 너무 많았어요!

깨끗하고 아나운서처럼 잘 읽는 목소리를 사람들이 좋아한다고 생각했는데…

오히려 기계 같고 인위적이라고 느끼는 것 같아요.

네이버 음성합성팀

Voice Maker 서비스란?

내 목소리와 CLOVA VOICE AI 기술이 만나,

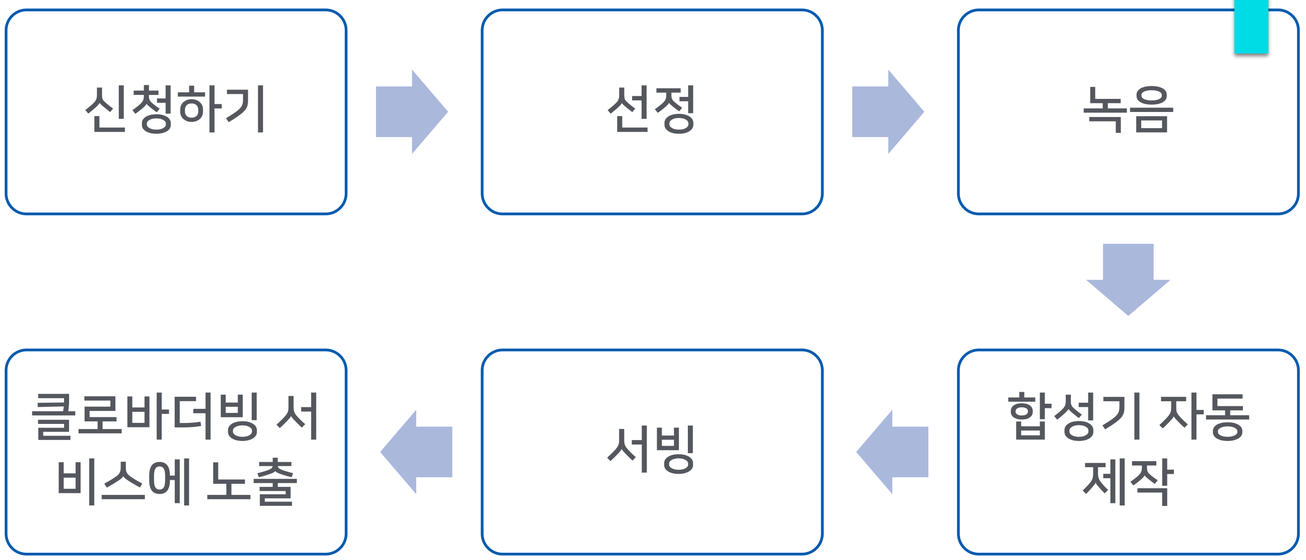
누구나

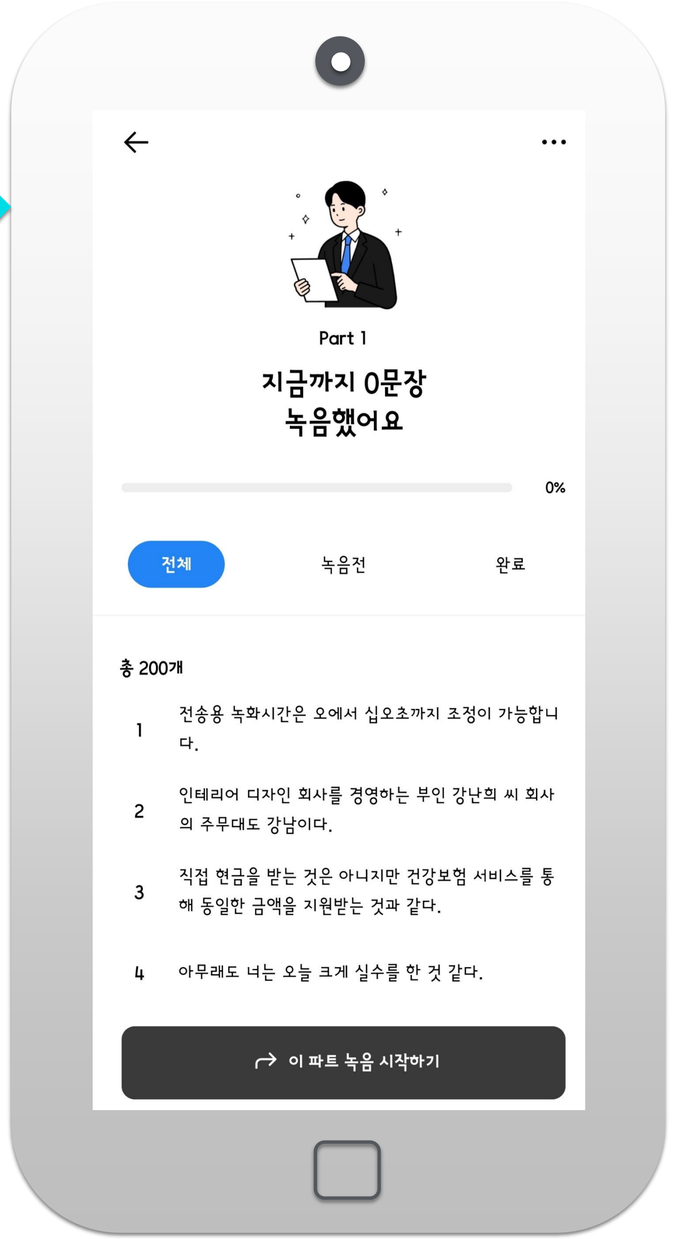
집에서

스마트폰 녹음으로

내 목소리 합성기를 만들어보자!

Voice Maker 서비스





5. Voice Maker 서비스를 위한 우리의 노력

스타일 특화 자체 스크립트 제작

DEVIEW 2021

- 챠프포프킨과 치스챠코프는 라흐마니노프의 피아노 콘체르토의 선율이 흐르는 영화.

- 증시 주변 여건이 조금씩 개선되면서 투신권이 모처럼 거래소 주식과 선물을 동시에 매수하였으나, 투자심리가 크게 위축된 일반 투자가들의 주식처분이 계속되면서 지수 낙폭이 깊어졌습니다.

한번 읽어볼까요?

스타일 특화 자체 스크립트 제작

DEVIEW 2021

합성기 스타일

기존	최근
낭독이나 안내용 사용처 예: 뉴스 본문 읽기	일상 대화, 어린아이, 구연 동화, 애니메이션 스타일 등등 사용처 예: 클로바더빙, 클로바램프

스타일 특화 자체 스크립트 제작

다양한 스타일 스크립트 제작

- 스타일별 맞춤 스크립트 제작
- 적절한 어휘 분포와 실제 사용빈도를 고려

쇼핑 호스트

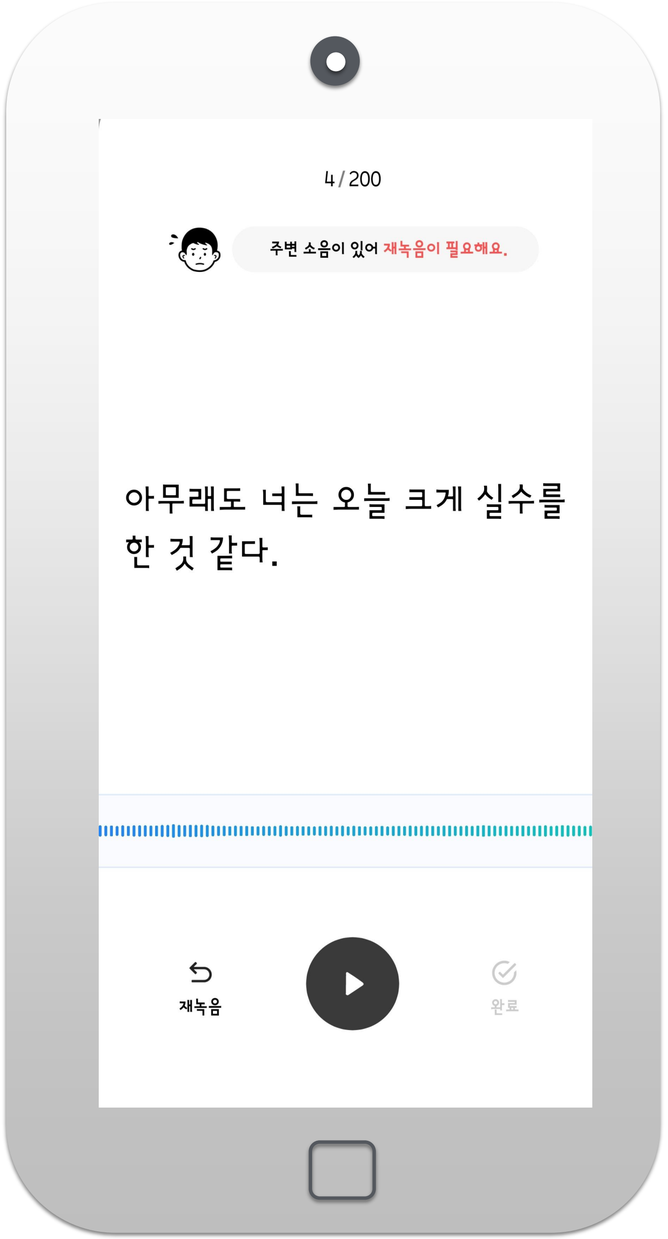
이렇게 놀라운 가격, 124만 9900원에 이 모든 것을 여러분들께 드립니다.

구연동화

달님도 별님도 크게 하품을 하고 있었어요.

녹음 품질 판단 알고리즘 개발

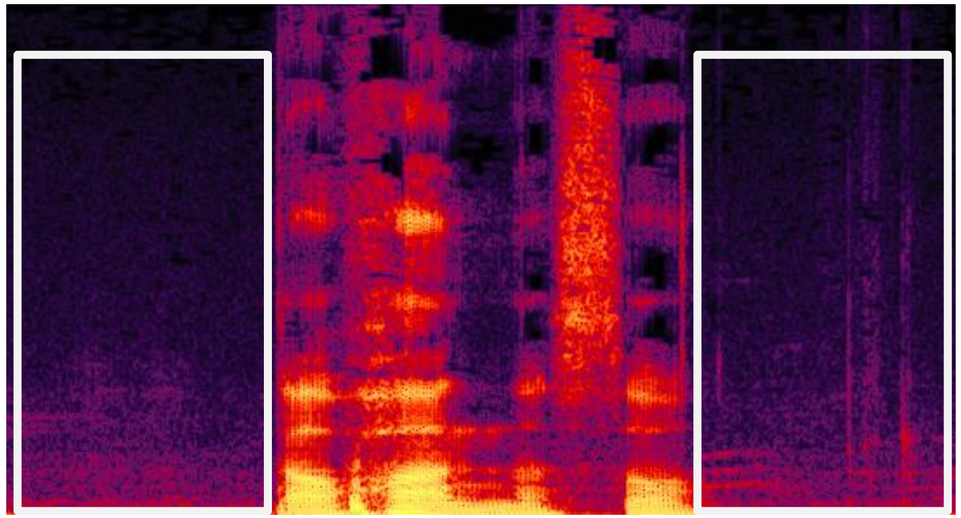
- 음성 대비 노이즈 큰 경우
- 비 음성 구간의 노이즈 레벨 판단
- 갑자기 책상 치는 소리(탁, 퍽) 처럼 목소리랑 특징이 확연히 다른 소리가 들어오는 경우



녹음 품질 판단 알고리즘 개발

DEVIEW 2021

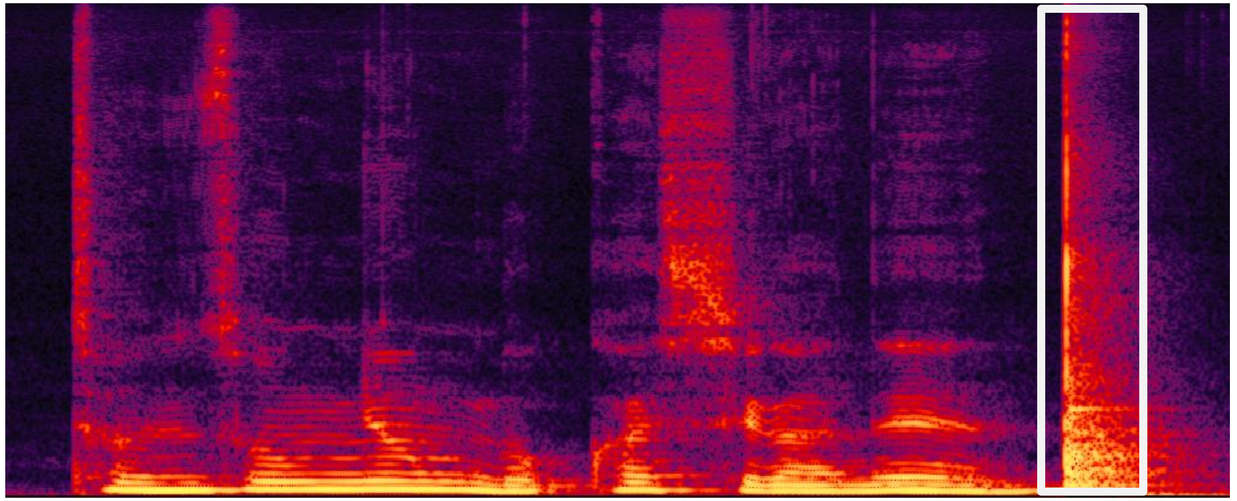
저품질 녹음 예시



녹음 품질 판단 알고리즘 개발

DEVIEW 2021

저품질 녹음 예시

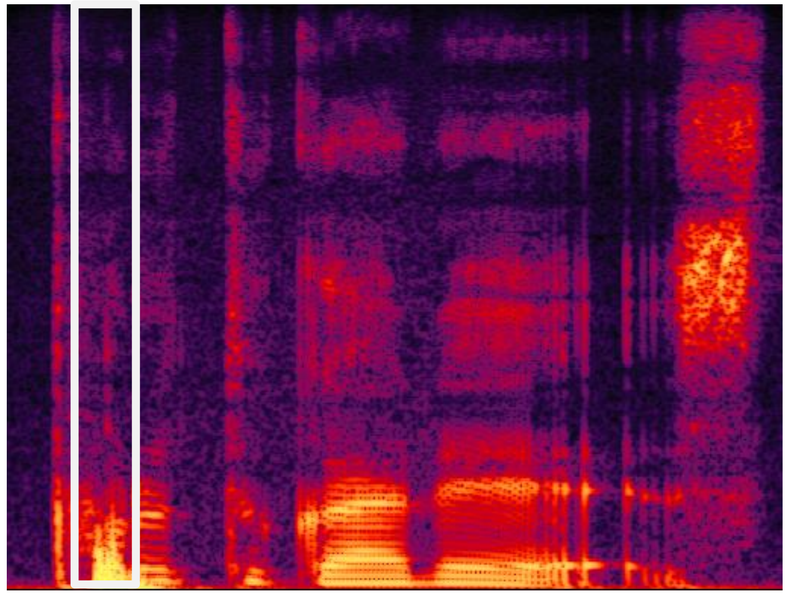


녹음 품질 판단 알고리즘 개발

DEVIEW 2021

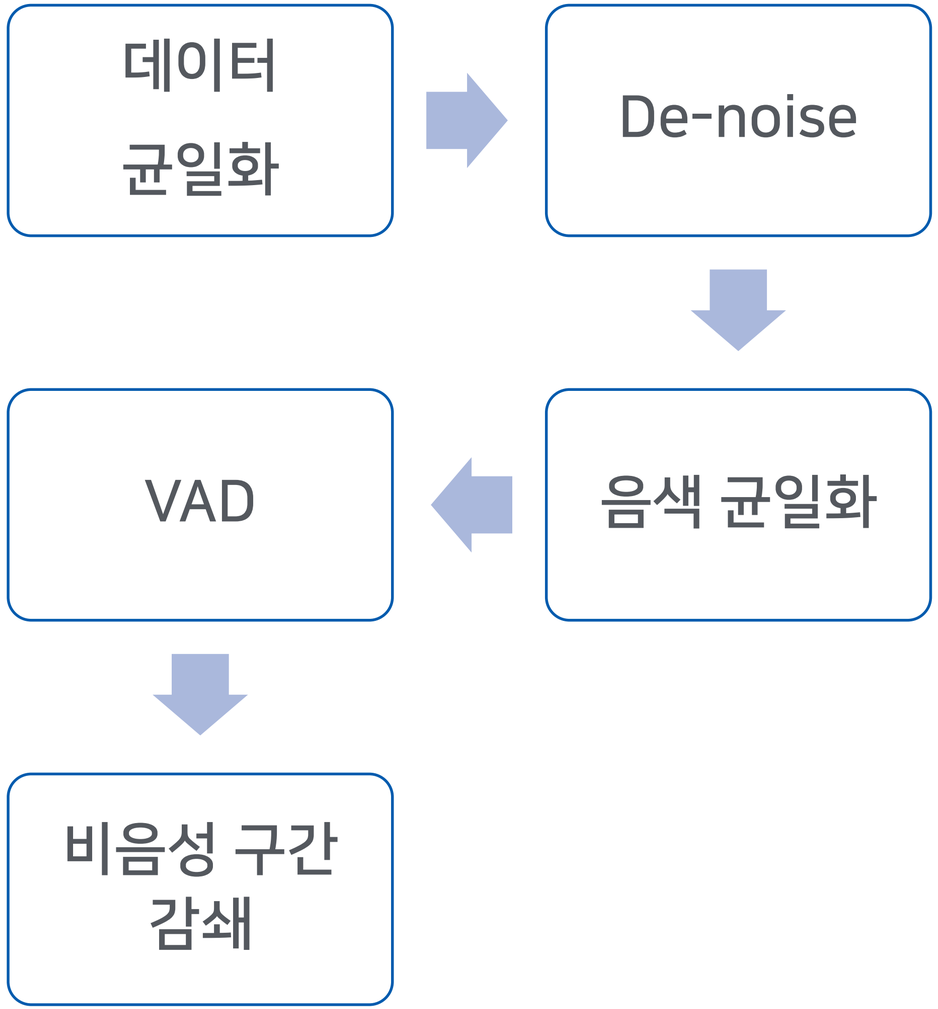
저품질 녹음 예시

센 파열음



모두 재녹음 필요로 판단!!

녹음 데이터 전처리 기술 고도화



데이터 균일화
- 파일 포멧, 볼륨, sampling rate

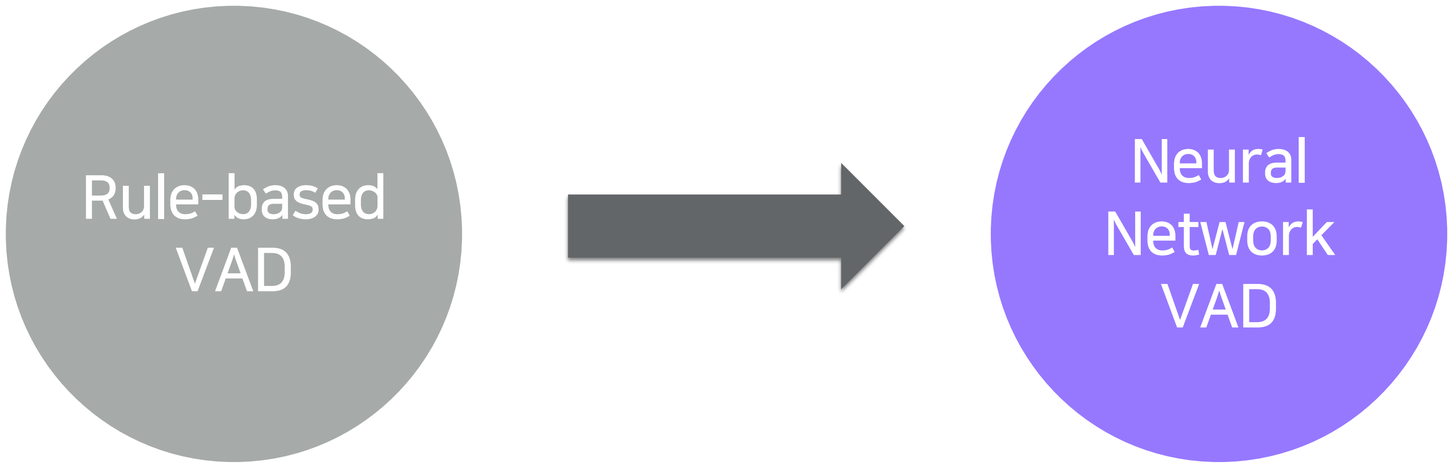
De-noise
- 배경 소음 제거

음색 균일화
- 기준이 되는 파일 자동 선택

VAD (Voice Activity Detection)
- 음성/비음성 구간 판단 후 비음성 구간 감쇄

녹음 데이터 전처리 기술 고도화

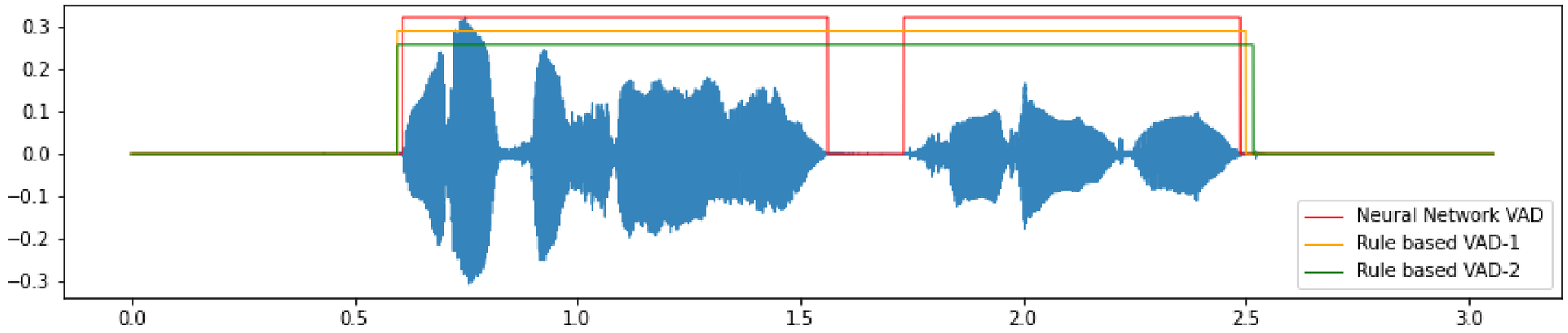
VAD (Voice Activity Detection)



녹음 데이터 전처리 기술 고도화

DEVIEW 2021

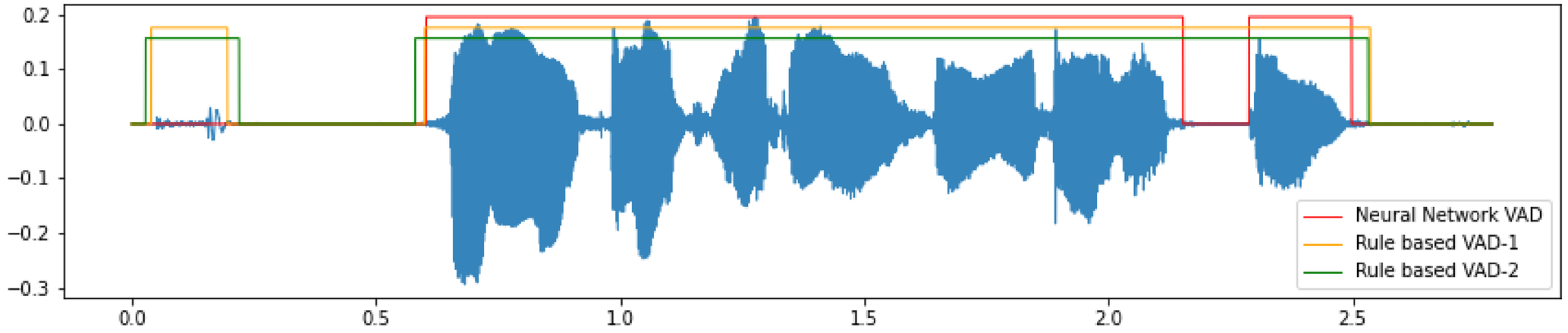
VAD별 결과



녹음 데이터 전처리 기술 고도화

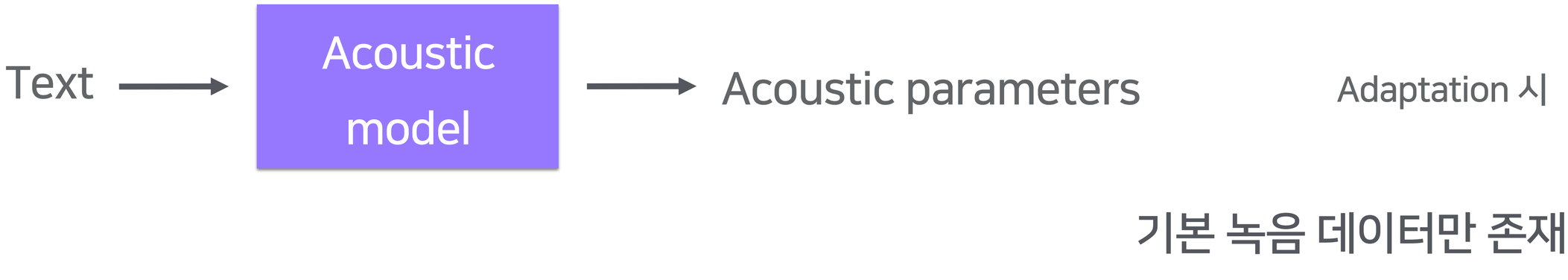
DEVIEW 2021

VAD별 결과



시작 부분에 배경 잡음

텍스트에 따라 스타일을 살리는 Acoustic Model

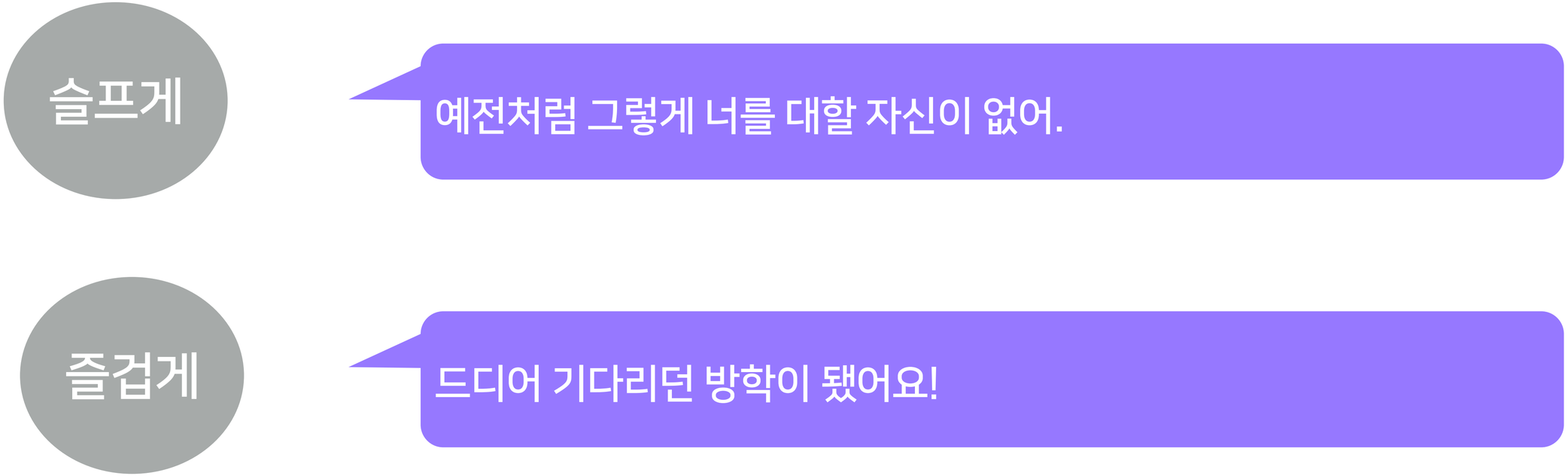


텍스트에 따라 스타일을 살리는 Acoustic Model

DEVIEW 2021

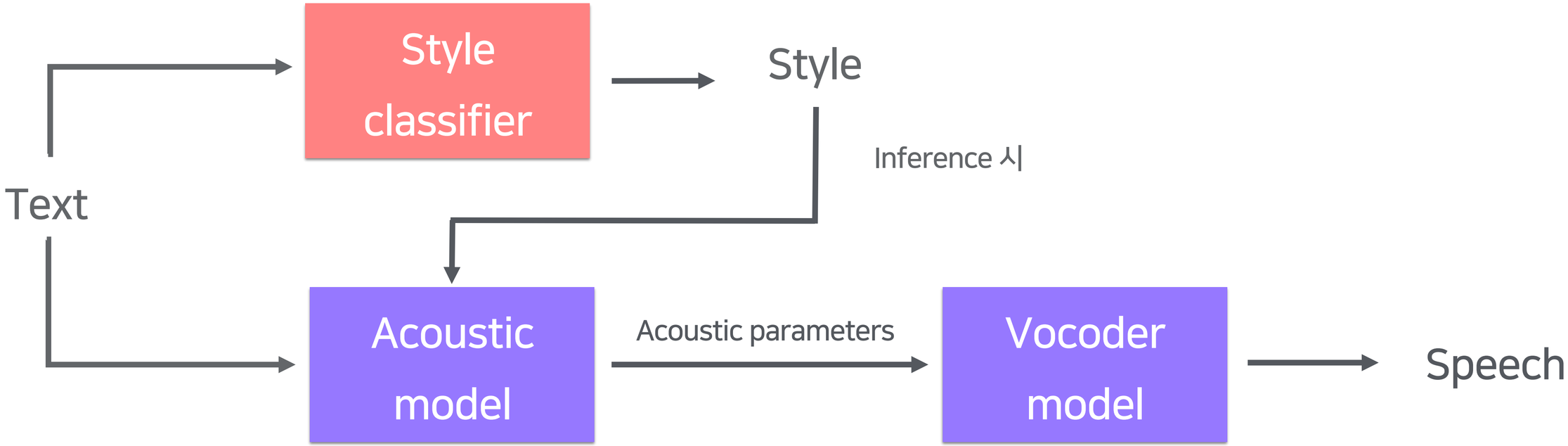
Style classifier

- 문장에 따라 어울리는 스타일을 자동 추정



텍스트에 따라 스타일을 살리는 Acoustic Model

Inference 시 텍스트에 따라 추정된 Style을 AM에 입력



텍스트에 따라 스타일을 살리는 Acoustic Model

DEVIEW 2021

합성음

슬프게

예전처럼 그렇게 너를 대할 자신이 없어.

즐겁게

드디어 기다리던 방학이 됐어요!

Acoustic Model with Duration

DEVIEW 2021

서비스의 요구사항

A 화자는 말하는 속도가 좀 느려요.

아바타와 연동하고 싶은데 시간 정보가 필요해요!

Acoustic Model with Duration

서비스의 요구사항

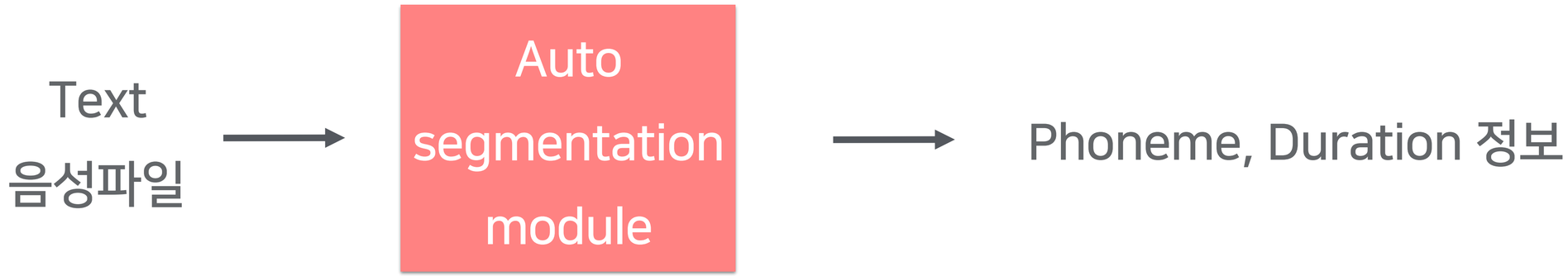
A 화자는 말하는 속도가 좀 느려요.

아바타와 연동하고 싶은데 시간 정보가 필요해요!

Duration 정보를 사용하는 모델 개발!!

Acoustic Model with Duration

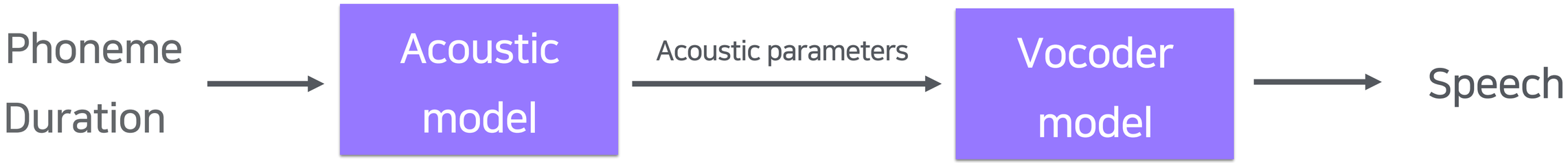
자동 전사 모듈 사용
- 수동 전사 대신 학습데이터를 위한 자동 전사



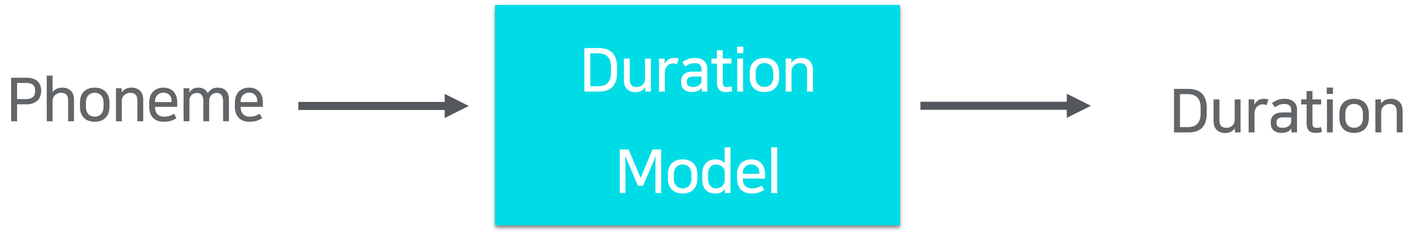
Acoustic Model with Duration

duration 정보를 사용하는 모델 개발

* Acoustic Model 학습 시



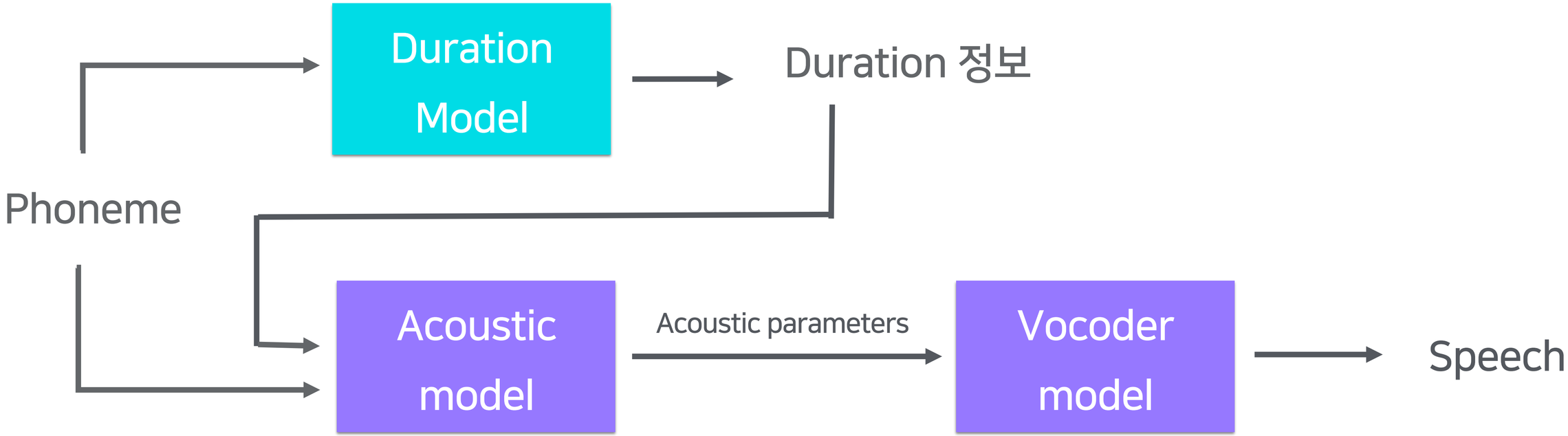
* Duration 모델 학습 시



Acoustic Model with Duration

DEVIEW 2021

Inference시



Acoustic Model with Duration

- 필요 시 자연스러운 합성음 속도 조절 가능

전체 문장의 특정 부분을 빠르게 말하고,
특정 부분은 느리게 말하는 것이 가능합니다.

duration x 0.7배

duration x 1.3배

Acoustic Model with Duration

- 합성음과 함께 Phoneme 별 Duration 정보 제공 가능
- 아바타와 연동도 가능해짐

알아서 검증하고 최적의 모델 선정

좋은 합성기 모델 판단 기준

loss가 낮은 모델?

발음이 또렷한 모델?

주관적 청취 평가로 듣기 좋은 모델 ?

알아서 검증하고 최적의 모델 선정

- 에러 유형 정의
- 수천의 검증 문장
- 매 epoch 마다 학습과 동시에 verification 수행
- 에러 개수와 loss 값들을 이용한 measure 개발

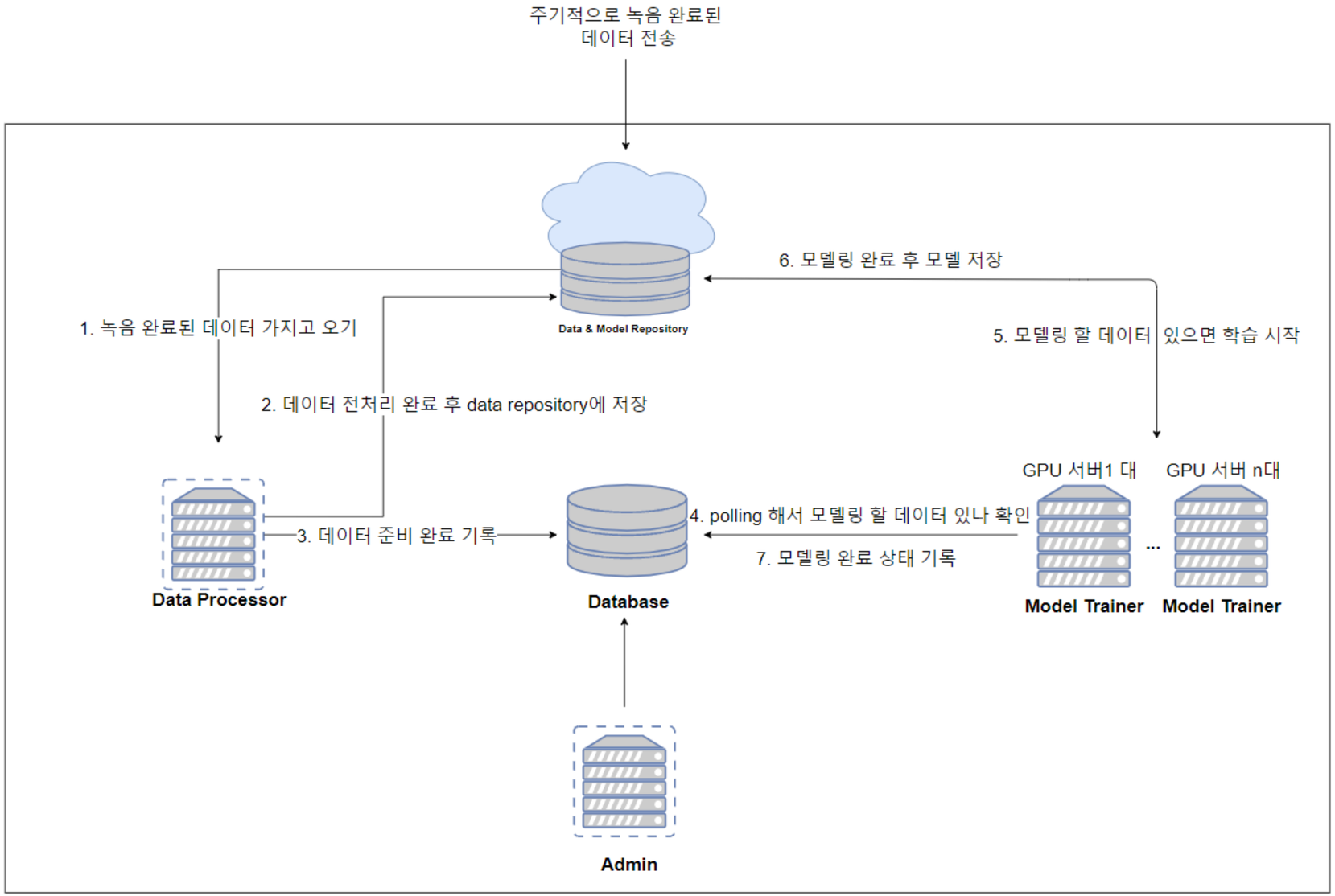
자동 모델 선택을 위해서는 이런 것들이 필요!

모델링 자동화 시스템 구축

한 땀 한 땀 장인의 손맛으로 만든 합성기

과연 많은 사람들의 목소리를 만들 수 있을까?

모델링 자동화 시스템 구축



6. 들어볼까요

개인화 합성기 시연

DEVIEW 2021

몰텐과 돈핀도 부부가 되어 새끼를 낳았습니다.

2018년 DEVIEW

130문장 -10분 데이터

2020년 DEVIEW

270문장 - 20분 데이터

2021년 DEVIEW

270문장 - 20분 데이터

휴대폰 녹음앱 사용, 모델링 자동화

개인화 합성기 시연

아무리 중요하다고 해도 한 가지만 보고 채용을 결정할 수 있을까요?

2018년 DEVIEW

130문장 -10분 데이터

2020년 DEVIEW

270문장 - 20분 데이터

2021년 DEVIEW

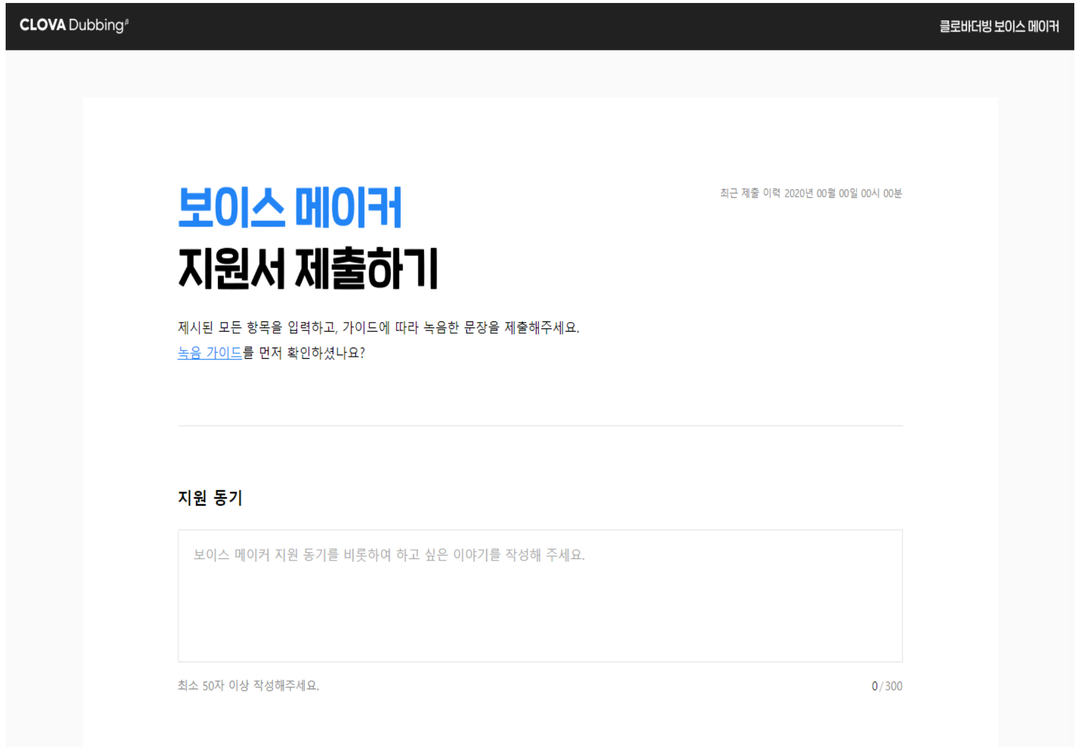
270문장 - 20분 데이터

휴대폰 녹음앱 사용, 모델링 자동화

7. 클로바 보이스의 주인공이 되려면 어떻게 하면 되나요?

클로바 보이스의 주인공이 되려면?

지원하기

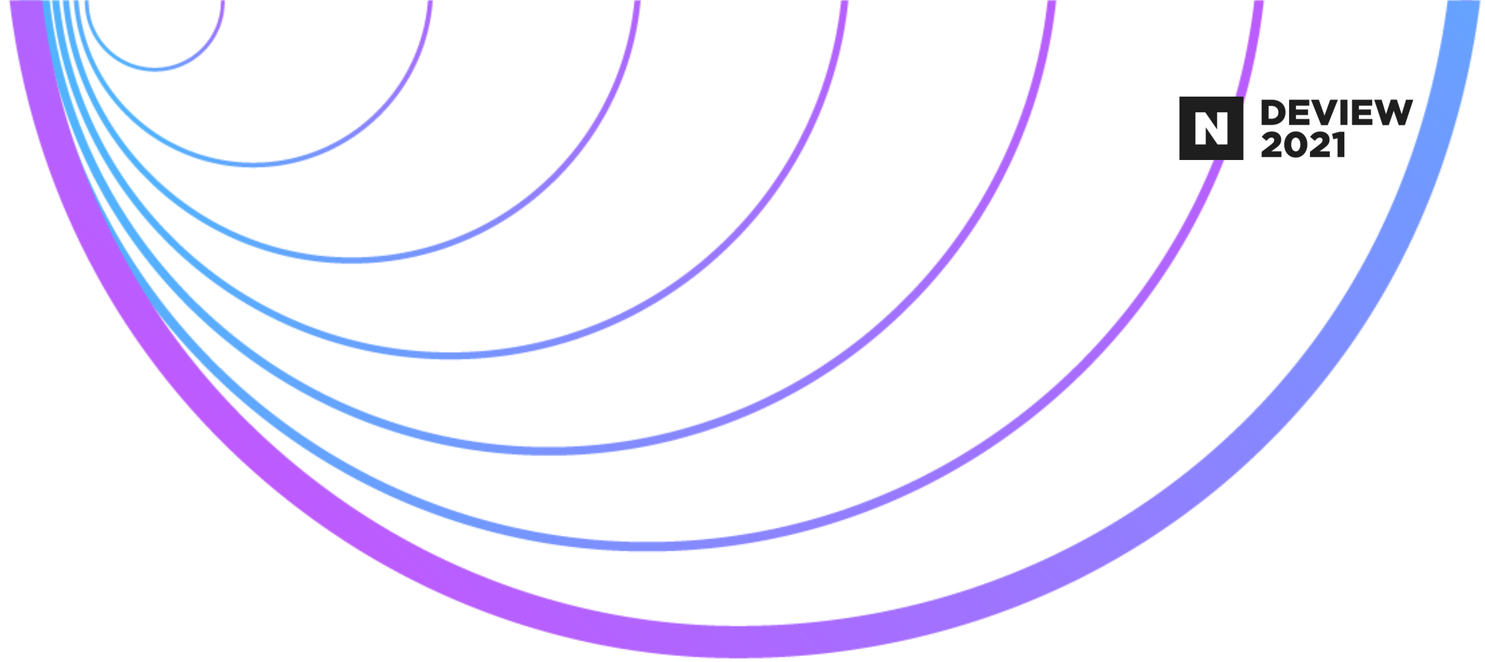


보이스 스타일 선택

AI 보이스로 만들고 싶은 스타일을 선택하고, 스타일을 최대한 살려 녹음한 파일을 제출해 주세요.
실제 녹음시, 아나운서와 유사한 기본 스타일은 약 1시간 분량, 나머지 스타일은 약 2시간 분량 녹음이 필요합니다.

녹음파일 제출

녹음 가이드에 따라 필수 문장 녹음이 정상적으로 완료되었는지 확인 후 제출해 주세요.
녹음 품질이 낮은 경우 선정 대상에서 제외될 수 있으니, 깨끗하게 녹음되었는지 확인해 주세요.



DEVIEW 2021

8. Next

Next

- 감정 표현 고도화
- Controllable TTS
  - 문장, 어절, 음절 단위로 속도, pitch, energy, 감정 등을 원하는 대로 조절
- Multilingual TTS
  - 한국어로만 녹음했지만 영어, 일본어도 할 수 있는 합성기

We are Hiring!!

DEVIEW 2021

여러분을 기다리고 있습니다!

>> 지원하기

- https://recruit.navercorp.com/naver/job/list/developer > CLOVA로 검색
- https://clova.ai/ko/research/careers.html

Thank You