| 2020- 2 Ybigta Conference

2020.
12

책 읽어주는 AI

ybigta

CONTENTS

01

Introduction

ybigta

01 | 1-1. 프로젝트 소개

안녕하세요! 여러분
EXID 하LI입니다
오늘 소개해드릴 내용은 바로 **음성합성**인데요
다들 전혀 눈치 채지 못하셨겠지만 놀랍게도
저의 목소리는 **딥러닝**으로 만들어 졌습니다

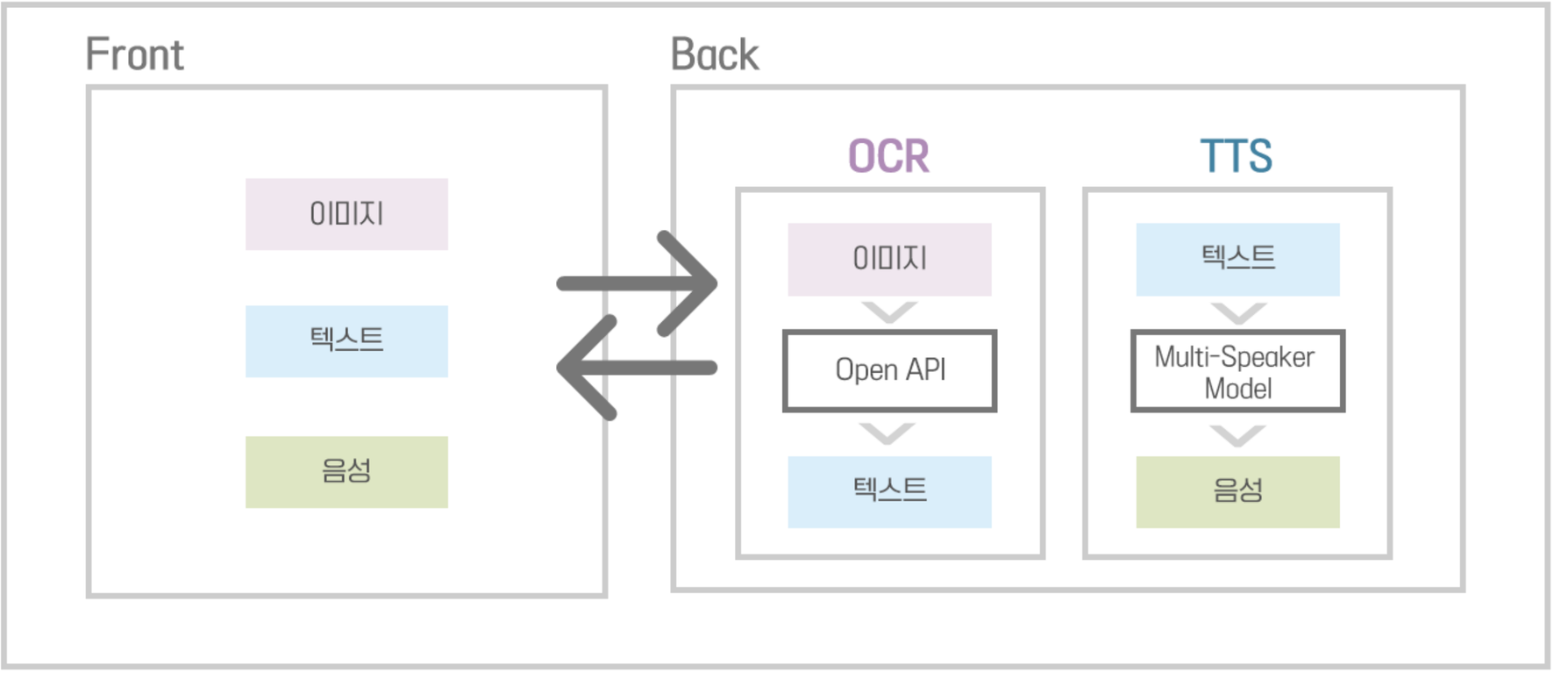
ybigta

01 | 1-1. 프로젝트 소개

" OCR + TTS + Application "

OCR을 통해 텍스트가 있는 이미지에서 텍스트를 추출하고 추출한 텍스트를 학습시킨 음성 합성 모델에 넣어 원하는 화자의 목소리로 합성. 이를 Application에서 돌아가도록 구현하는 것이 프로젝트의 목적!

Application



02

Dataset and Training Model

ybigta

02 | 2-1. 데이터 설명 : 데이터 수집

학습을 위해 기본적으로 오디오와 텍스트의 pair가 필요하다.

텍스트를 input으로 집어넣었을 때 원하는 목소리의 오디오를 얻기 위해서
오디오와 이에 대응하는 텍스트 데이터가 필요하다.
그렇다면 어떠한 오디오 데이터가 학습에 적절할까? 다음과 같은 기준으로 선정하였다.

노이즈를 줄임으로써
학습이 잘 될 수 있도록

01. 음질이 깨끗하고 배경음이 거의 없음

02. 화자가 한 명이어야 함

학습에 필수 x
프로젝트를 위해서

03. 들었을 때 화자를 분별 가능 (유명인)

ybigta

02 | 2-1. 데이터 설명 : 데이터 수집

01. 오디오 데이터 수집

손석희

손석희 50h: 앵커브리핑(약 950개)

http://news.jtbc.joins.com/article/article.aspx?news_id=NB11754993
뉴스 아이디에 따라 950개 리스트가 있고 ts파일(동영상) 크롤링
-> wav로 변환해서 오디오 수집

KSS Dataset

KSS 12h : Korean TTS를 위해 전문 여성 성우로 녹음된 데이터셋

https://www.kaggle.com/bryanpark/korean-single-speaker-speech-dataset
기본 오디오 타입이 wav로 되어있음

EXID 하니

하니 3h : '내 이름은 삐삐 롱스타킹' 오디오 북

https://audioclip.naver.com/audiobooks/901DFE68BCt
직접 녹음을 통해 wav로 저장

ybigta

02 | 2-1. 데이터 설명 : 데이터 수집

01. 오디오 데이터 수집

박근혜 6h+: 유튜브 영상 크롤링 (14개)

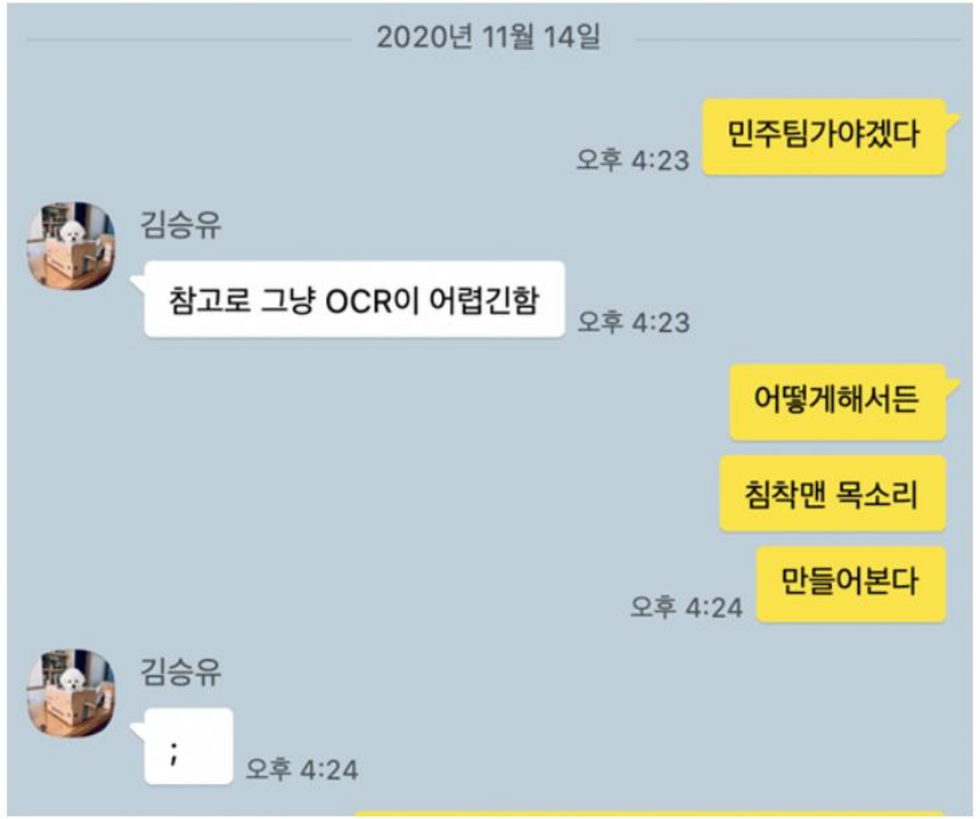
https://www.youtube.com/watch?v=5G4o-v8QfFw&ab_channel=YTNnews
14개 리스트가 있고 ts파일(동영상) 크롤링
>> wav로 변환해서 오디오 수집

침착맨 4h+ : 침착맨 유튜브 영상 (13개)

https://www.youtube.com/watchv=Ui6VVxN9w3A&ab_channel=%EC%B9%A8%EC%B0%A9%EB%A7%A8
13개 링크 리스트가 있고 mp4 파일(동영상) 크롤링
>> wav로 변환해서 오디오 수집

02 | 2-1. 데이터 설명 : 데이터 수집

01. 오디오 데이터 수집



하지만..... 결과는 과연?

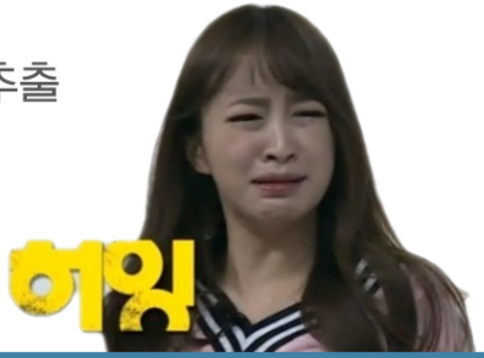
ybigta

02 | 2-1. 데이터 설명 : 데이터 수집

02. 텍스트 데이터 수집

- kss는 텍스트가 기본적으로 제공
- 나머지 데이터는 **음성인식 API**를 사용해서 텍스트를 추출
- API는 **구글 STT API**랑 **ETRI Open API**를 사용

  * 하니는 오디오[북]임에도 스크립트가 없음..



Google STT API

Speech To Text

https://cloud.google.com/speech-to-text/docs/?hl=ko

ETRI STT API

API 호출 1일 허용량

기술명	API명	1일 허용량
음성인식 기술	▪ 한국어 인식 API ▪ 중국어 인식 API ▪ 독어 인식 API ▪ 스페인어 인식 API ▪ 베트남어 인식 API ▪ 영어 인식 API ▪ 일본어 인식 API ▪ 불어 인식 API ▪ 러시아어 인식 API	1,000건/일 (60초 이내/건당)

Etri: https://aiopen.etri.re.kr/guide_recognition.php

ybigta

02 | 2-1. 데이터 설명 : 데이터 전처리

01. 오디오 전처리

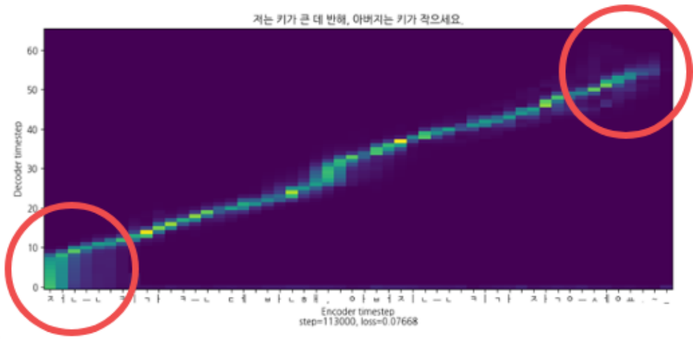
1 오디오를 공백 기준으로 분할

12초 이하의 오디오를 인풋으로 받는 모델이고, 문장 단위가 아닌 공백 기준으로 나누는 이유는 **공백이 불규칙적으로 있으면 띄어쓰기를 잘 학습하지 못하기 때문**

* 공백 기준으로 나누는 데에는 크기 기준과 길이 기준이 있음. **일정 크기 이하의 소리가 일정 시간 지속되면 공백으로 판단**하고 분할한다

2 앞뒤 공백 제거

앞뒤 공백이 있는 경우 텍스트와 매칭이 되지 않는 문제가 있어서 librosa.effects.trim()으로 제거하였음 그림에서 보면 어텐션이 양끝이 잘 안잡힘



3 음량 정규화

오디오 별로 소리의 크기가 다른데 이를 그대로 두면 소리의 크기도 변수가 되기 때문에 정규화

ybigta

02 | 2-1. 데이터 설명 : 데이터 전처리

02. 텍스트 전처리

하니, 박근혜, 이말년

음성인식 후 하나하나 노가다로 오탈자 수정

손석희

- 데이터가 많아 하나하나 수작업 불가
- 음성인식 후 영어는 jtbc -> 제이티비씨, 단위 km -> 킬로미터
  숫자 -> 하나 둘 셋 넷… 로 변환하고 자리수에 맞추어 다시 변환
- 이후 한글 이외에 언어가 있으면 제외(약 2/3 정도 남음)

최종적으로

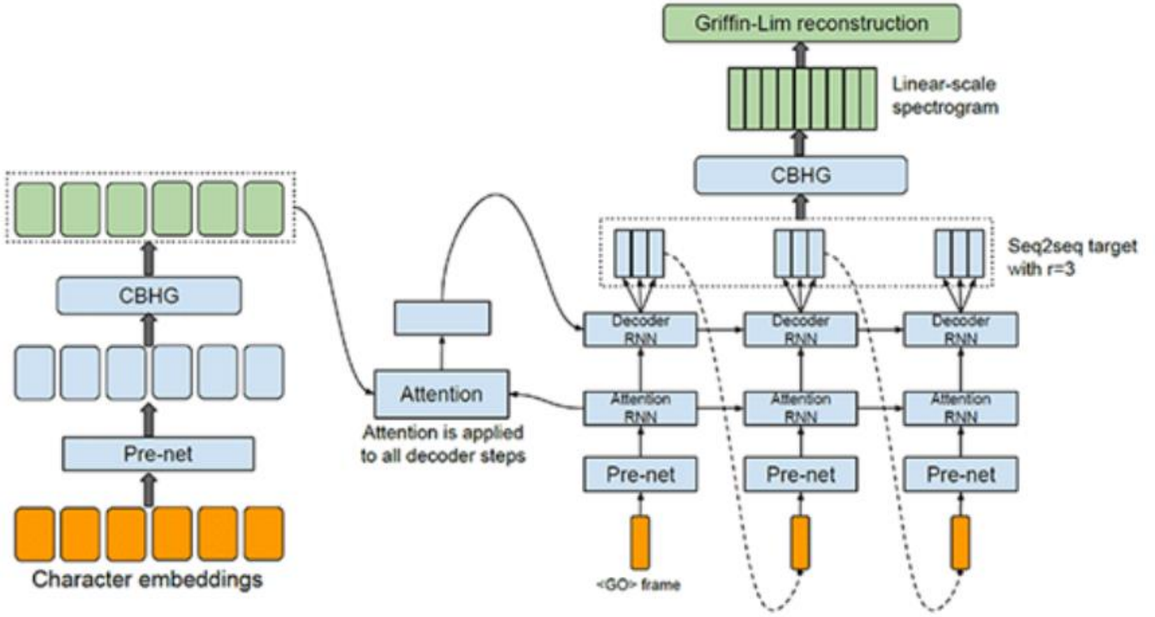
- 텍스트의 길이가 비슷하도록 **min-max 필터링**
- 앞에서 전처리한 오디오와 텍스트를 묶어서 **하나의 넘파이 파일**을 생성

02 | 2-2. 모델 설명

Multi-Speaker Tacotron: Tacotron+deepvoice

기본적으로 Wang et al.(2017)이 제안한 Tacotron의 구조를 따름.

Tacotron은 입력 문자열에서 스펙트로그램을 출력하는
attention 메커니즘 기반의 순환신경망(RNN, recurrent neuralnetwork) 인코더-디코더와
음성 합성부(보코더)로 이루어져 있다



ybigta

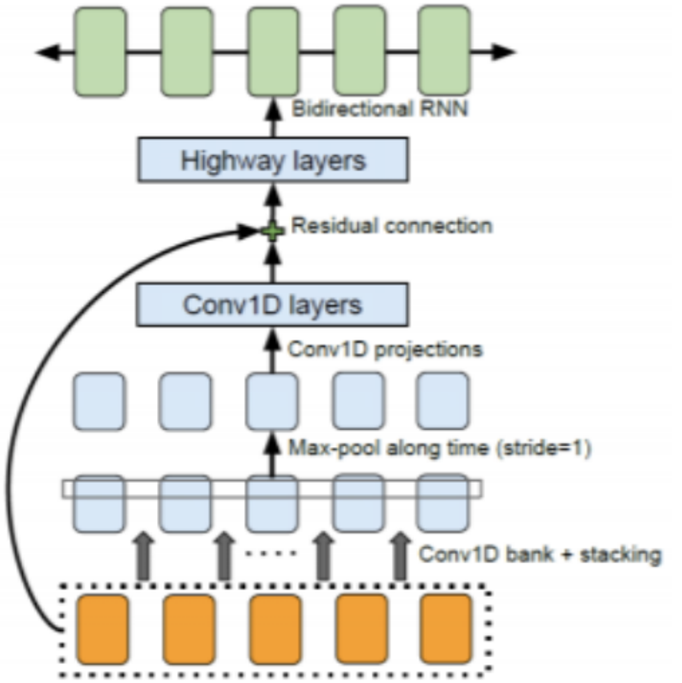
02 | 2-2. 모델 설명

Tacotron

Tacotron은 입력 문자열에서 스펙트로그램을 출력하는
attention 메커니즘 기반의 순환신경망(RNN, recurrent neuralnetwork) 인코더-디코더와
음성 합성부(보코더)로 이루어져 있다

인코더

인코더는 입력 문자 임베딩 열을 받아 **annotation vector를 출력** 하는 부분으로서,
보다 robust한 인코더를 구현하기 위해
RNN이 아닌 **CBHG**를 사용



* **CBHG 모듈**은 Lee et al.(2016)이 제안한 기
계 번역을 위한 인코더로부터 착안된 구조로,
1D convolution bank, highway 네트워크, bidirectional
gated recurrent unit(GRU)로 구성

ybigta

02 | 2-2. 모델 설명

Tacotron

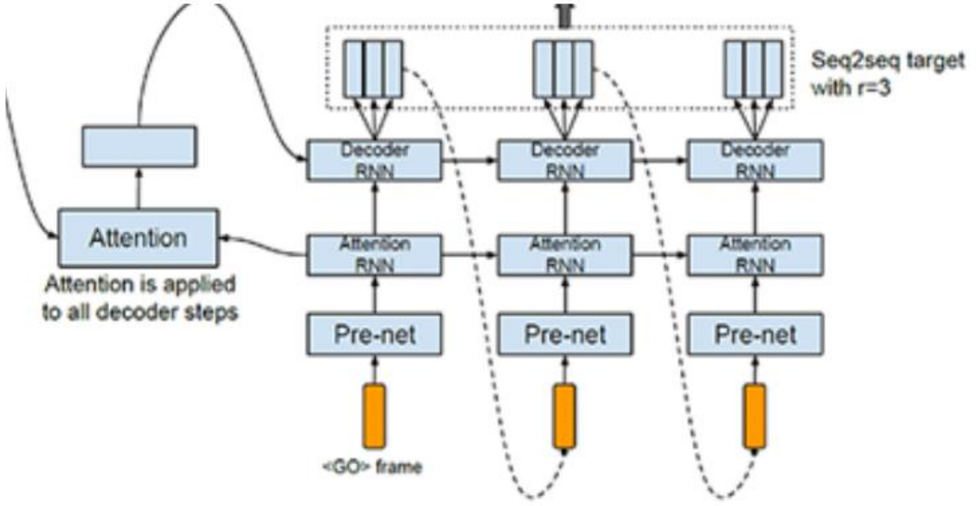
Tacotron은 **입력 문자열에서 스펙트로그램을 출력하는**
attention 메커니즘 기반의 순환신경망(RNN, recurrent neuralnetwork) **인코더-디코더**와
음성 합성부(보코더)로 이루어져 있다

디코더

디코더는 **특정 time step** 프레임의 **스펙트로그램**을 입력으로 받고, **다음 time step** 프레임의 **스펙트로그램**을 출력한다.

최종적으로 디코더는 **Mel-Spectrogram**을 출력한다.

디코더 구조는 모델마다 조금씩 상이한데 오른쪽과 같이 attention RNN을 따로 두는 방식도 있다.



ybigta

02 | 2-2. 모델 설명

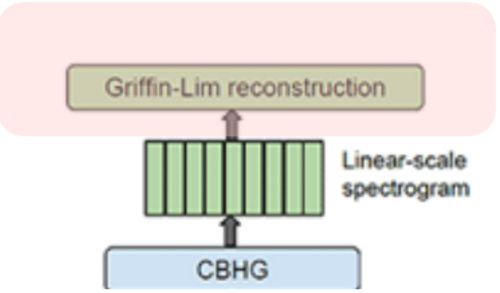
Tacotron

Tacotron은 **입력 문자열에서 스펙트로그램을 출력하는**
attention 메커니즘 기반의 순환신경망(RNN, recurrent neuralnetwork) 인코더-디코더와
음성 합성부(보코더)로 이루어져 있다

보코더

디코더에서 생성된 Mel-Spectrogram은 PostNet에 해당하는 **CBHG**를 거치게 되며 최종적으로 **Linear-Spectrogram**을 생성하게 된다.

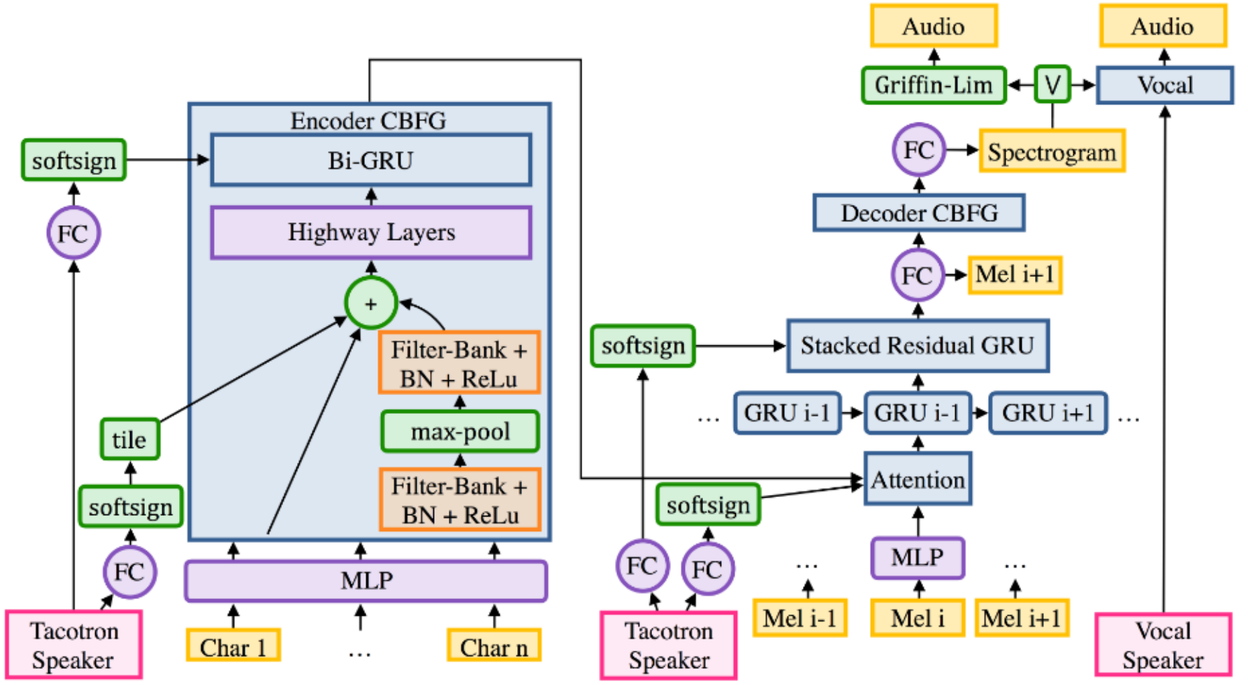
Linear-Spectrogram은 **Griffin-Lim Algorithm**이라는 음성 재구성 알고리즘을 통해서 **음성으로 변환**되게 되는데 이 알고리즘이 바로 **보코더**에 해당한다.



02 | 2-2. 모델 설명

Multi-Speaker Tacotron: Tacotron+deepvoice

다시 Multi-Speaker Tacotron으로 돌아오면,
Multi-Speaker Tacotron은 Tactron과 DeepVoice 모델을 합한 모델로,
단일 화자 뿐아니라 복수의 화자에 대한 학습을 가능하게 한다



Deep Voice 2: Multi-Speaker Neural Text-to-Speechhttps://arxiv.org/abs/1705.08947
Tacotron: Towards End-to-End Speech Synthesishttps://arxiv.org/abs/1703.10135

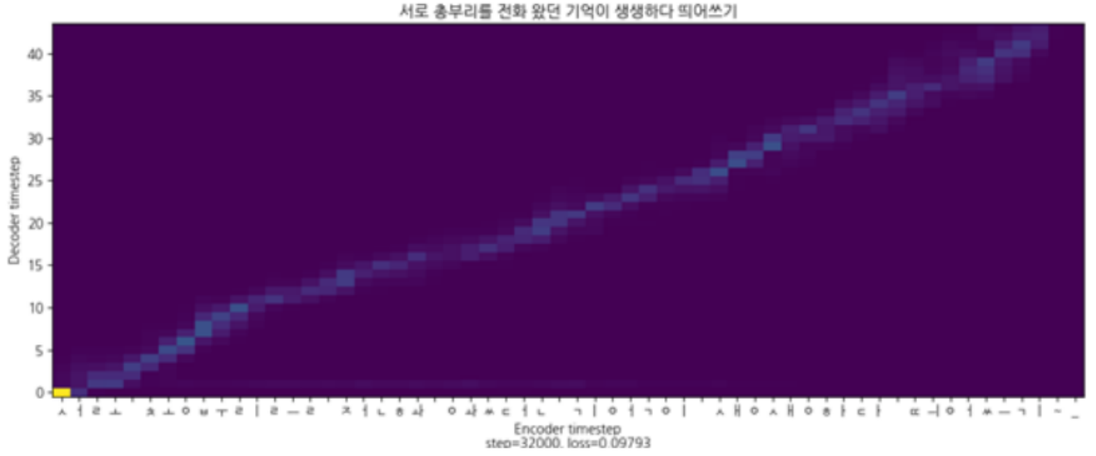
02 | 2-3. 모델 튜닝 및 결과 : 모델에 따른 결과

Single Model vs Multi Model

화자에 따라 말하는 스타일은 다르긴 하지만, 한국어의 공통적인 요소를 학습하게 되는 멀티 모델이 싱글 보다 학습이 안정적으로 되어 어텐션이 잘 잡힌다.

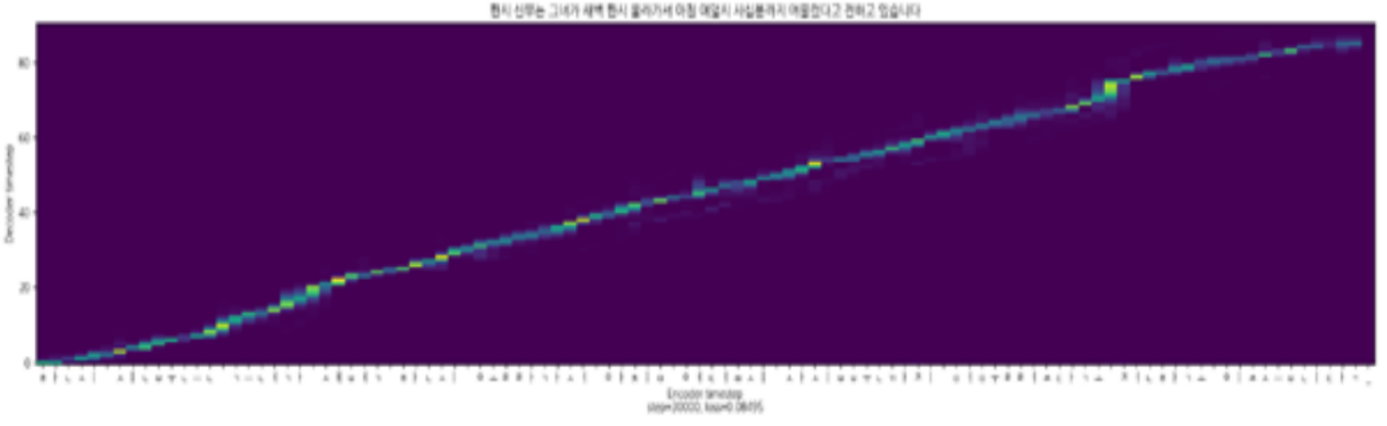
Single Model

손석희



Multi Model

손석희+하니+KSS

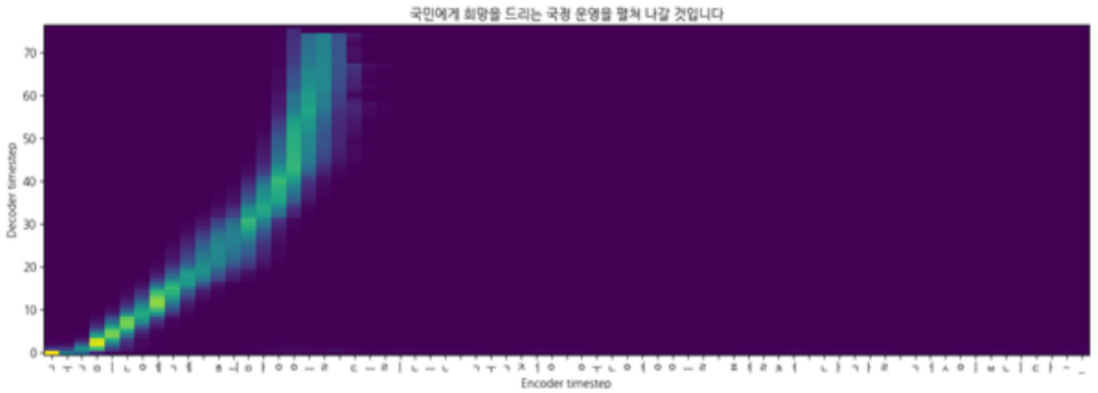


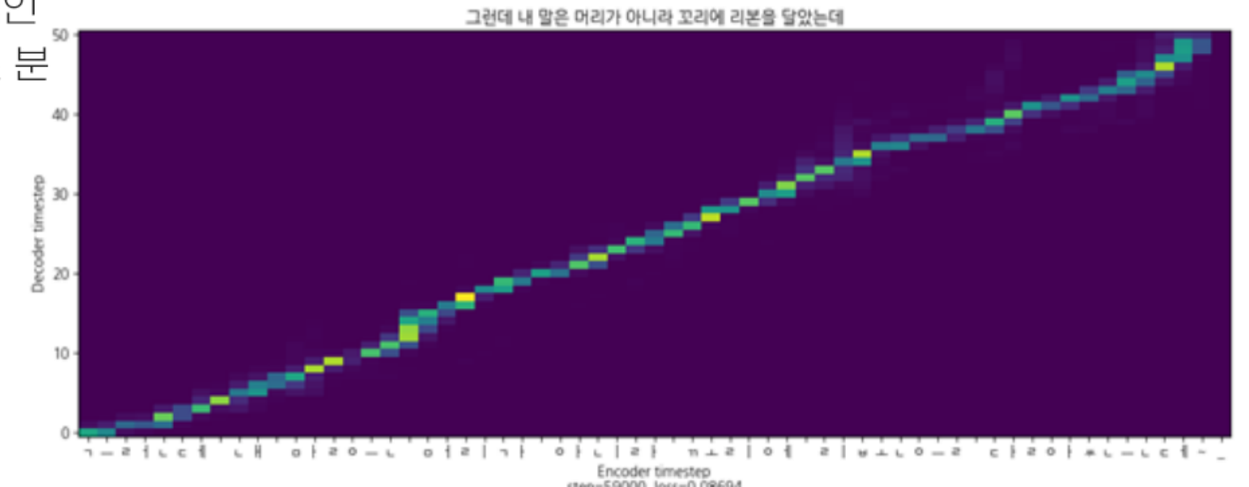
02 | 2-3. 모델 튜닝 및 결과 : 데이터에 따른 결과

무조건 다양한 화자 데이터를 사용?

멀티 모델의 성능이 좋은 걸 이제 알겠음.
그런데 무조건 다양한 화자의 데이터를 학습시키는 것이 좋을까?

위는 5명의 화자로 학습시킨 것
아래는 3명의 화자로 학습시킨 것(59k step)
그런데 오히려 화자가 3명일 때의 어텐션이 더 잘 잡히는 것을 볼 수 있다.





이는 이말년 데이터가 다른 데이터와는 달리 불규칙적인 발화 스타일을 가지고 있어서 노이즈로 작용한 것으로 분석된다.

02 | 2-3. 모델 튜닝 및 결과 : Reduction factor(하이퍼파라미터)에 따른 결과

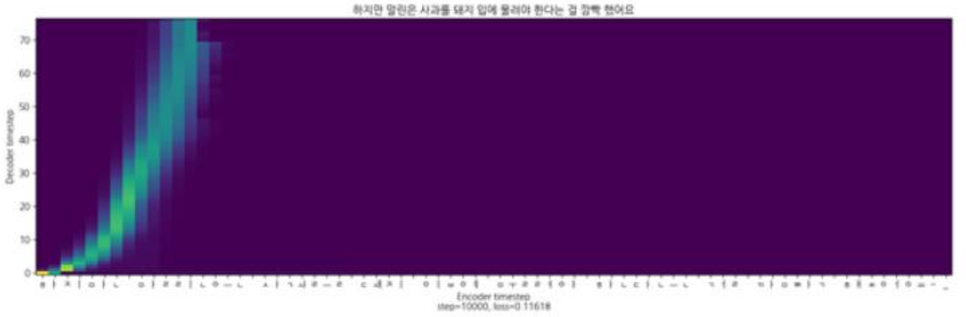
하이퍼 파라미터가 몇일 때 성능이 좋을까?

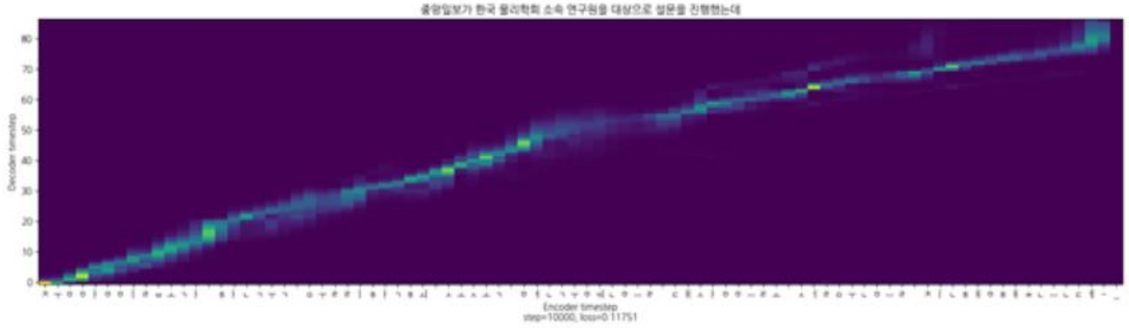
디코더의 time step 당 예측하는 프레임의 개수를 reduction factor(r)라고 부른다. 선행연구에 따르면 r이 4-10일 때 작동함을 확인되었다. 선행연구에 따르면 5일 때가 일반적으로는 청취 성능이 더 좋았으나, 문장에 따른 기복이 더 심하여 평가할 때는 4를 택하였다.

하지만, 우리의 결과는 5가 더 좋았다

동일한 10k 스텝에서 위가 reduction factor = 4, 아래가 5인데 위는 어텐션이 잘못된 방향으로 되고 있다. 이러면 아무리 학습을 잘하더라도 뒤에 텍스트에 대해서는 말을 할 수가 없게 된다..

반면 아래는 아직 수렴하지는 않았지만 어텐션의 방향은 잘 잡아가고 있다

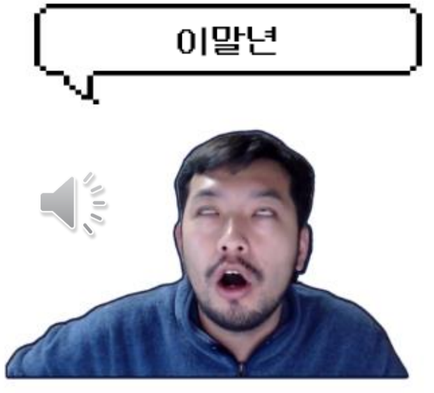




02 | 2-3. 모델 튜닝 및 결과 : 화자별 결과

화자별 TTS 결과

"흔들리는 꽃들 속에서 네 샴푸 향이 느껴진 거야"

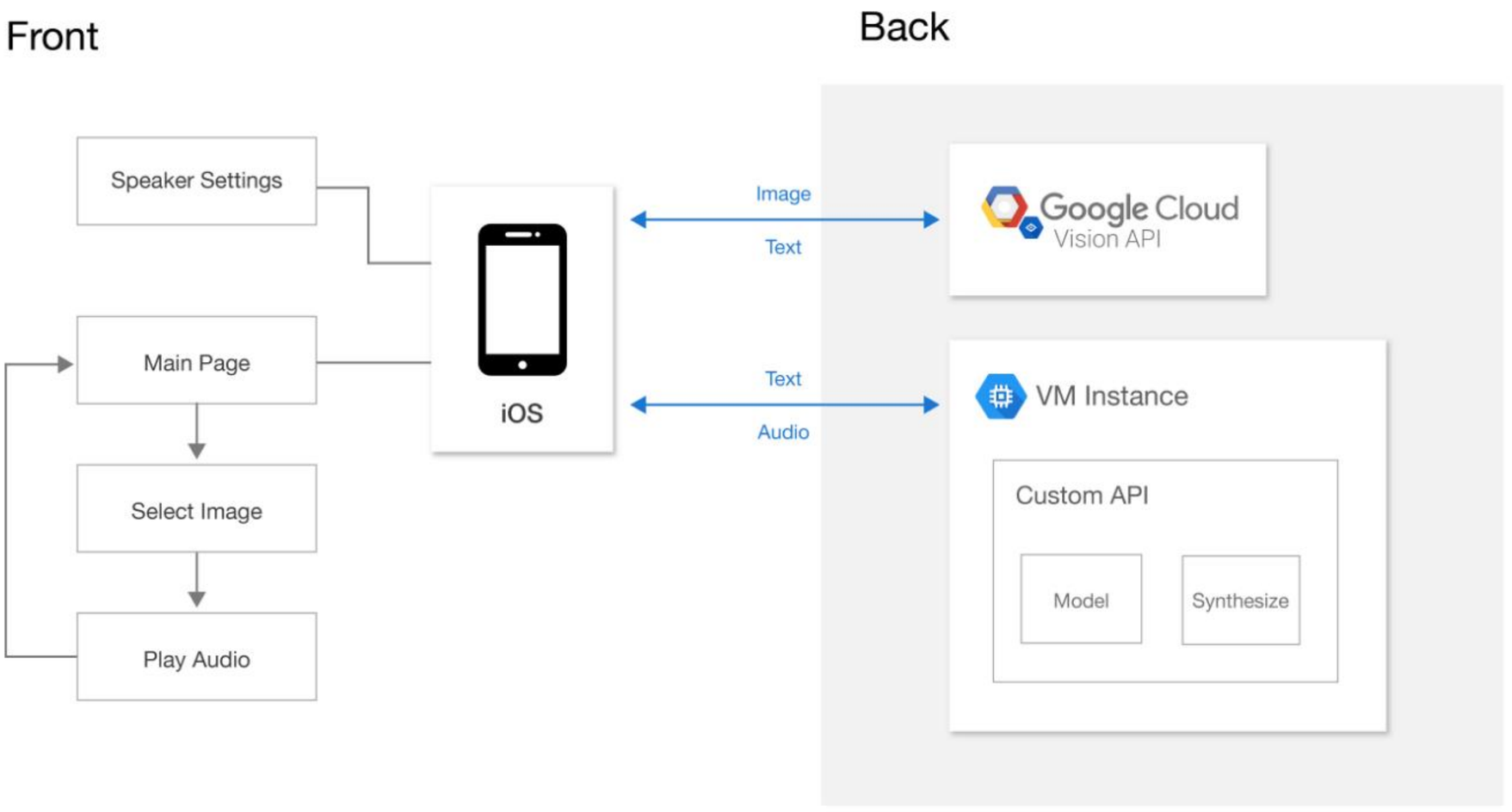


03

Application

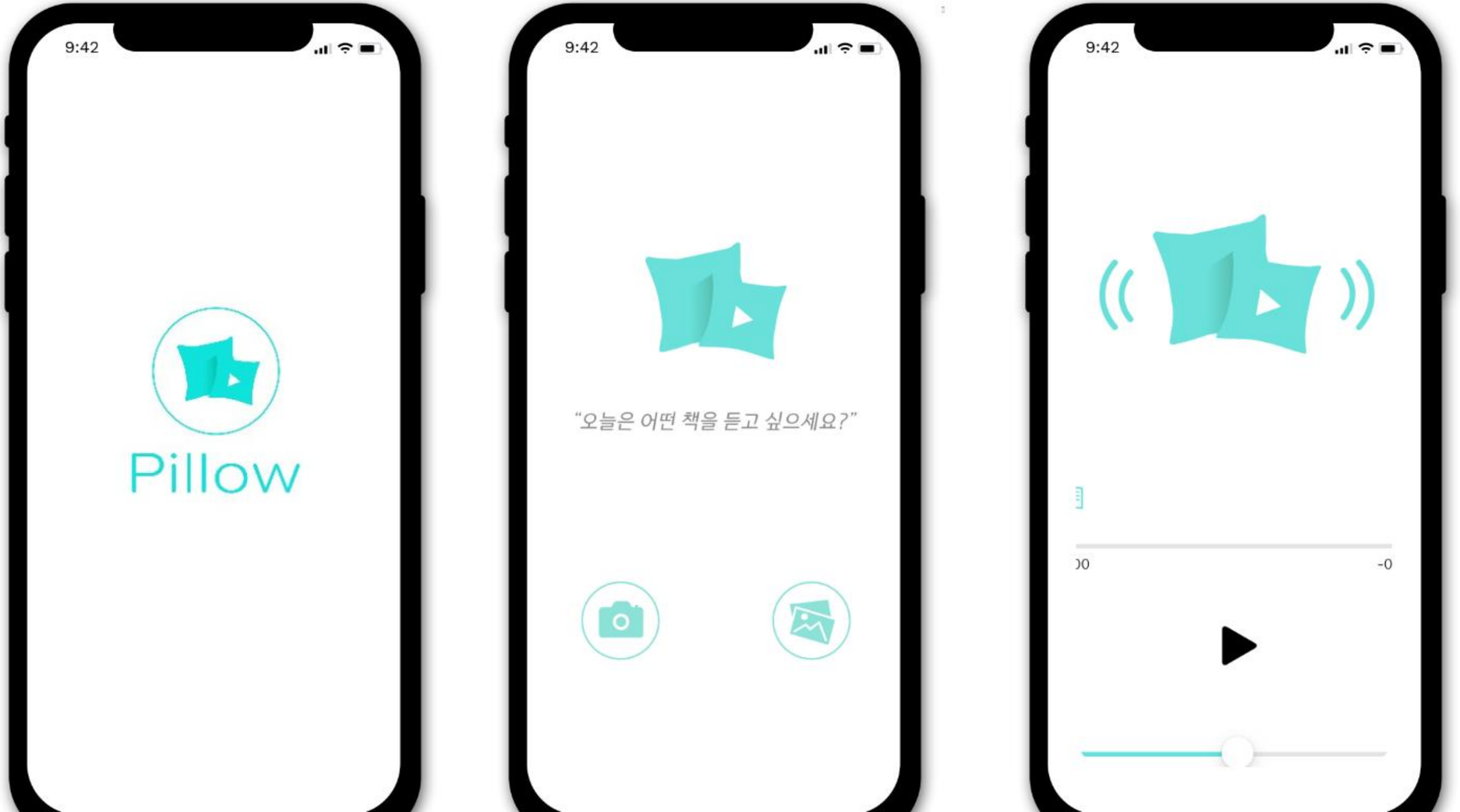
ybigta

03 | 3-1. 앱 구현 : Swift를 이용한 iOS 앱 개발



ybigta

03 | 3-2. 결과 시연



Thank you very much
Happy New Year!