Dalarm

달달한 콤퓨터 달콤팀

김세희 강승군 남병욱 위광진

목차

01 프로젝트 소개

프로젝트 소개

알람

유명인의 목소리를 이용

1. 사용자가 입력한 메세지를
   유명인의 목소리로 듣는 알람 어플리케이션
2. TTS(Text-to-Speech)를 위해서는
   유명인 목소리를 이용한 딥러닝 과정이 필요



02 수행과정

수행과정

1. 어플리케이션 구현

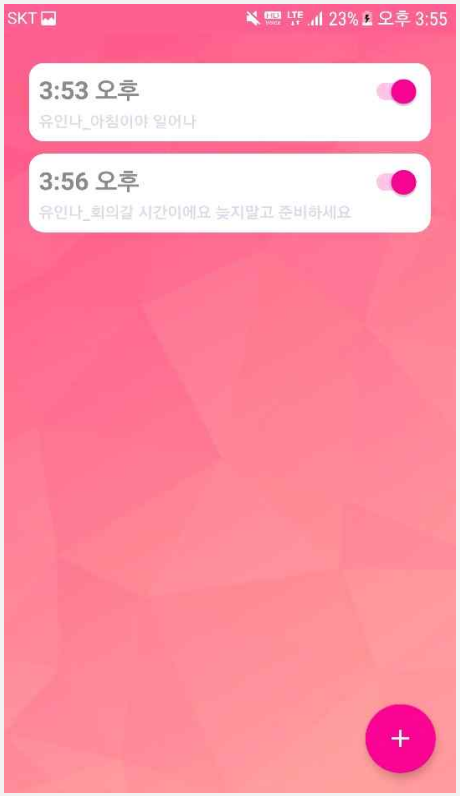
2. 딥러닝 진행

3. 데이터셋 수집 및 학습

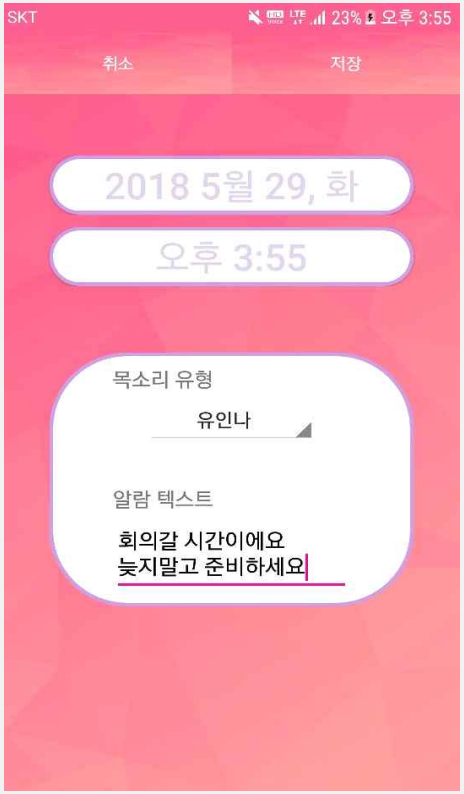
수행과정

어플리케이션 구현

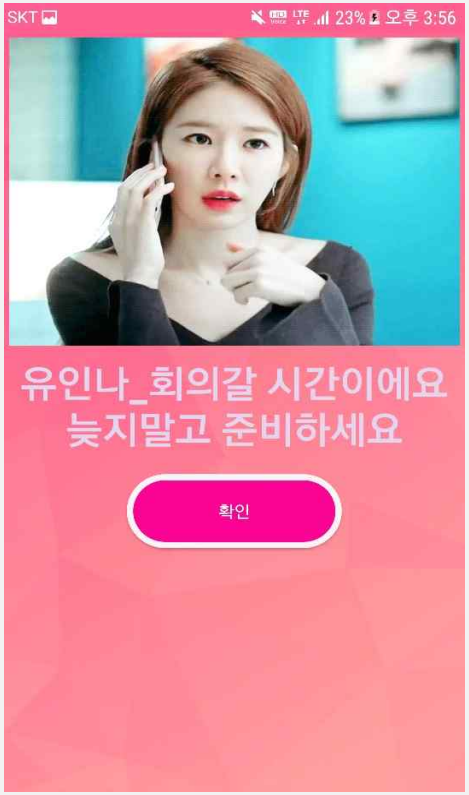
어플 로딩화면



생성된 알람 리스트 화면



알람 설정시 화면



알람 받을때 화면

수행과정

딥러닝 진행

- 뒤의 '딥러닝 학습' 페이지에서 저희 팀이 쓰는 '타코트론'에 대한 알고리즘 설명과 함께 실제 수행과정을 자세히 보여드리겠습니다.

수행과정

데이터셋 수집 및 학습

- 먼저, 손석희의 데이터셋은 이미 공개되어 있는 것을 바탕으로 타코트론 알고리즘의 하이퍼 파라미터를 변경하며 딥러닝 최적화를 진행하였습니다.

- 다음은 공개되어 있는 유인나의 라디오 음성데이터를 받아 Google speech API를 이용해 1차적인 텍스트틀을 만든 후 팀원 넷이서 직접 음성파일을 듣고 텍스트를 수정하는 작업을 통해 음성파일과 텍스트의 매칭률을 높여 최종 딥러닝 성공률을 높였습니다.

- 당초 프로젝트 기획 당시 유인나와 손석희 이외에도 인물을 추가할 수 있다면 하기로 계획했었고 따라서 라디오를 진행하고 있는 연예인 박명수를 선정하여 직접 데이터셋을 모으는 것부터 딥러닝까지의 단계를 직접 진행시켰습니다.

수행과정

데이터셋 수집 및 학습

- 박명수의 데이터셋은 각 라디오 방송국에서 운영하고 있는 '팟캐스트'의 RSS주소를 통해 충분한 음성파일을 확보한 다음, 1개당 1시간 가량의 음성파일을 타코트론 알고리즘을 이용하여 각각 '한 문장' 단위의 음성파일들로 나누어줍니다. 이후 Google speech API를 사용하여 딥러닝의 입력값으로 쓰일 '음성파일 이름-텍스트'의 틀을 만들고 팀원 넷이서 직접 음성파일을 들으며 텍스트와 매칭하고 수정하는 작업을 통하여 정확도를 크게 향상시켰습니다.

수행과정

데이터셋 수집 및 학습

- 또한 손석희의 딥러닝 학습 과정에서 학습 완료의 결과 음성에서 노이즈가 섞이는 것을 확인하고 이를 없애기 위한 작업을 진행하였습니다. 타코트론은 본래 학습이 잘 완성된 하나의 목소리 모델에 학습되지 않은 다른 목소리를 덧붙여 러닝시키면 후자의 목소리 모델을 더 완성도 있게 잘 만들어낼 수 있다는 특성이 있습니다. 따라서 달콤팀의 팀원 2명이 각각 직접 10시간 이상의 음성녹음을 통해 고품질의 데이터셋을 만들고 이를 바탕으로 손석희, 유인나, 박명수의 목소리를 합성하여 보다 고품질의 목소리를 완성시켰습니다.

03 딥러닝 학습

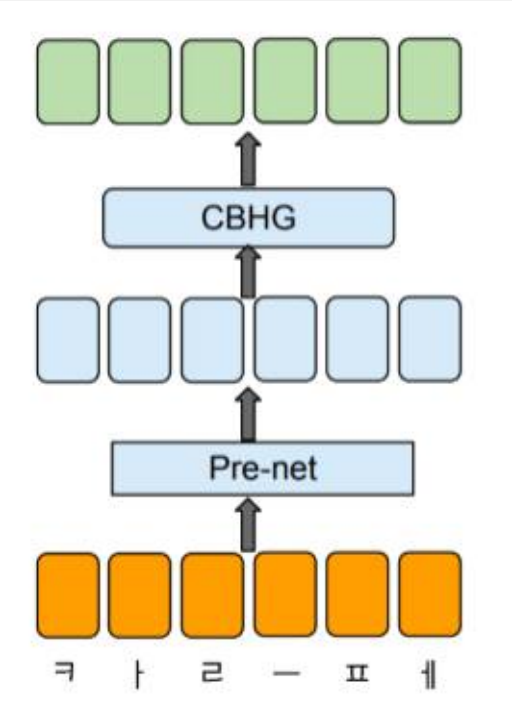
딥러닝 학습

사용자가 입력한 메세지를 유명인의 목소리로 들려주는 TTS(Text-to-Speech)를 위해서는 유명인의 목소리를 딥러닝 시켜주는 과정이 필요합니다. 따라서 목소리와 그에 대응되는 글자를 대량으로 입력시킨 후 우리가 사용할 알고리즘 '타코트론'을 이용하여 딥러닝 학습을 진행합니다. 이 학습이 본 달람 프로젝트의 핵심이자 근본적인 챌린지입니다.

음성 합성 Process

딥러닝 학습

Encoder



Encoder의 인풋값은 학습을 시킬 음성과 텍스트의 pair입니다.

첫 번째로 Encoder에서는 인풋값으로 들어온 문장을 자음모음으로 분리시켜서 각각을 정수값을 가지는 벡터로 character embedding 을 시킵니다.

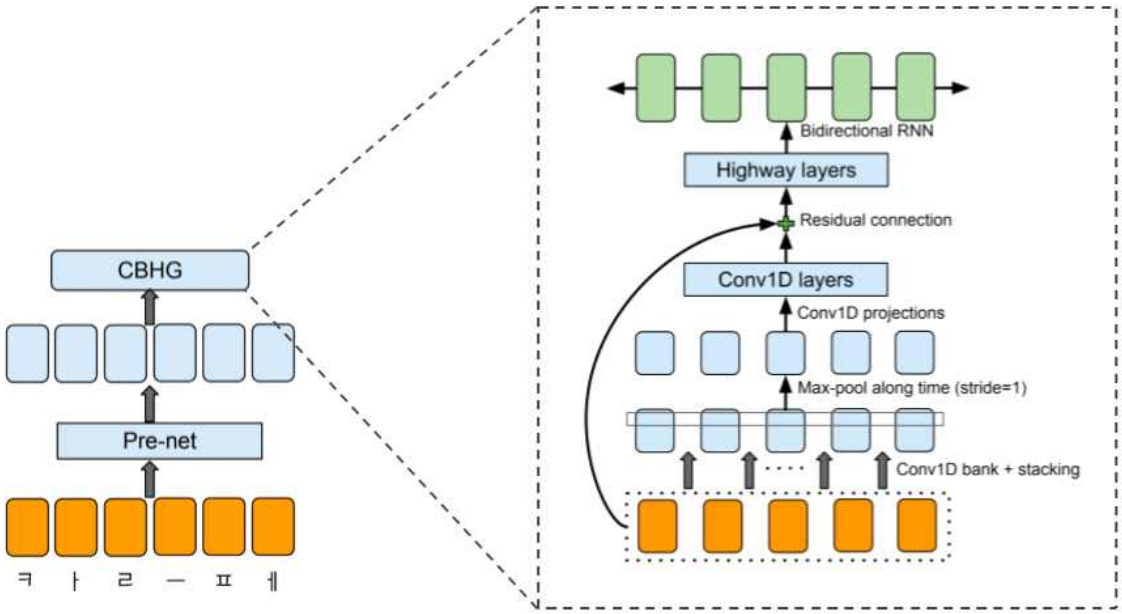
여기서 character embedding이란 자음 모음으로 나눈 글자를 딥러닝학습을 위해 숫자로 변환하는 작업을 말합니다.

두 번째로 pre-net 부분에서는 character embedding 한 입력값을 가지고 reLu activation 함수를 사용하여 히든 레이어를 2번 거쳐 새로운 embedding을 출력합니다.

세 번째로 CBHG 모듈에서는 pre-net의 아웃풋 값을 입력 받습니다.

딥러닝 학습

CBHG 모듈



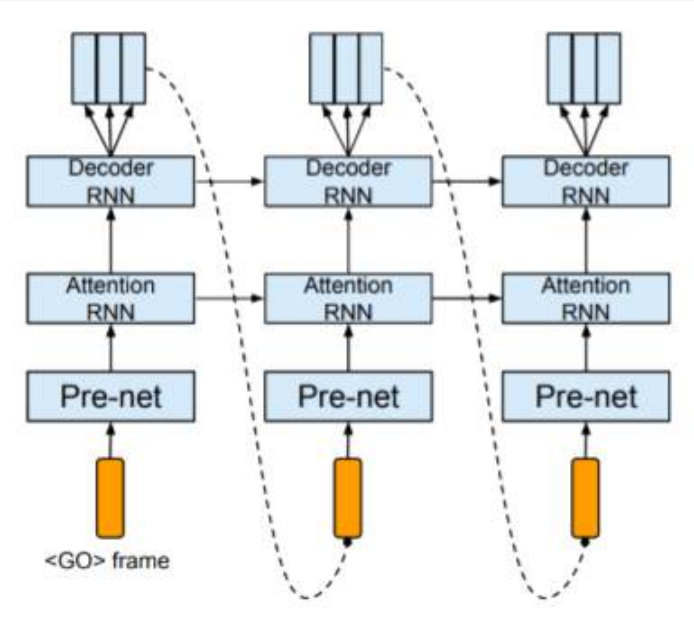
CBHG에서는 conv계산, max pooling, projection, residual connection과 같은 딥러닝
과정을 통해 highway layers로 구성된 highway network를 구성합니다.

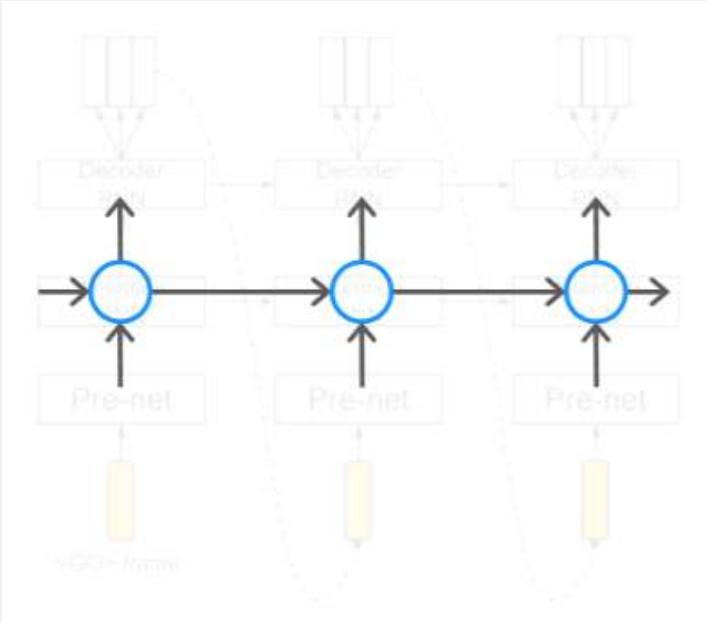
Highway network를 통해 문장에 대한 high level features를 뽑아낼 수 있고,

이렇게 구성된 highway network를 텍스트, 음성의 시퀀셜 피쳐를 추출하기 위해
bidirectional GRU RNN에 쌓아줍니다.

딥러닝 학습

Decoder





제 1 Decoder 단계는 모든 값이 0으로
초기화된 <GO> 프레임부터 시작합니다.

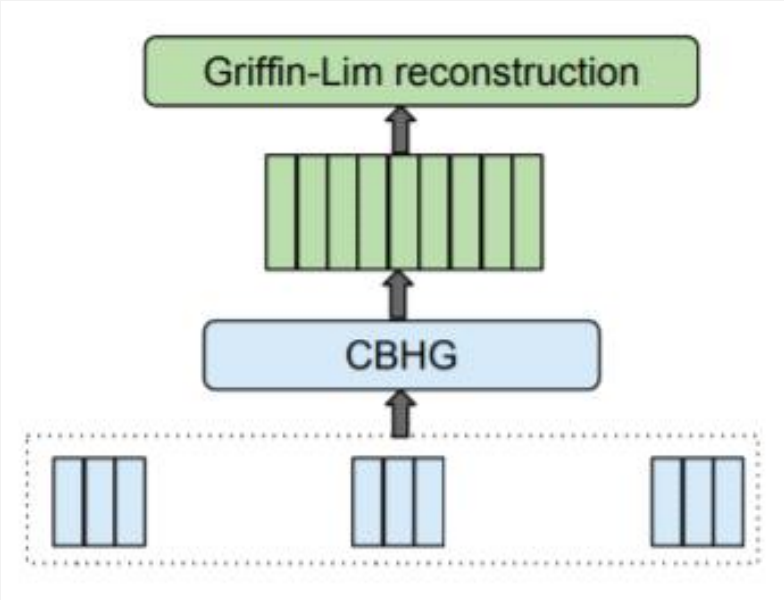
Pre-net과 Attention RNN을 거치고
Encoder output을 받아온
Decoder RNN을 거쳐서
첫 번째 스펙트로그램을 만들어냅니다.

여기서 스펙트로그램이란 간단하게
음성으로 변환되기 전 숫자 값들 이라고
할 수 있습니다.

이 후에는 Decoder 단계 t에서 r개의
예측 최종 프레임은 단계 t+1의 디코더에
입력으로 공급되며 입력 프레임은
첫 번째 단계와 같은 과정을 거쳐
스펙트로그램을 만듭니다.

딥러닝 학습

Vocoder

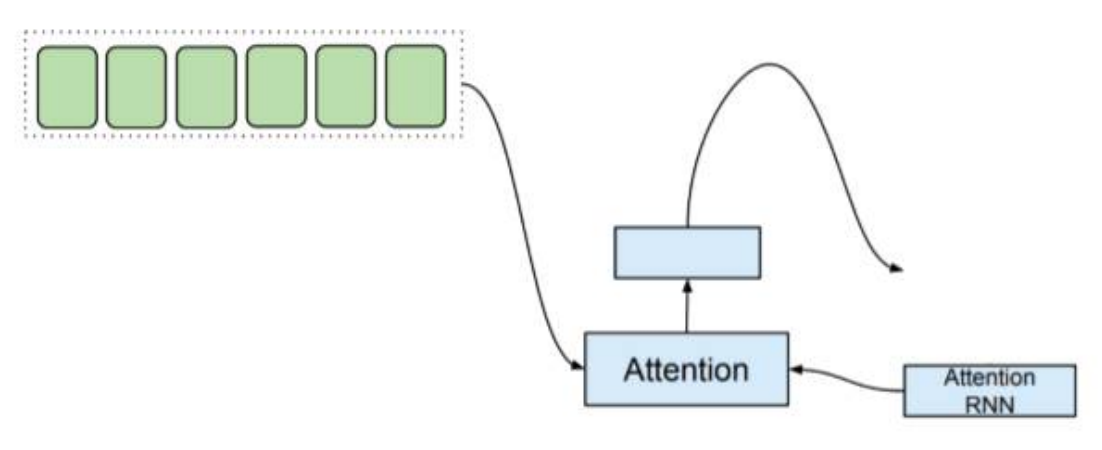


Vocoder는 Decoder에서 만들어진
스펙트로그램을 인풋값으로 받아서

CBHG와 Griffin-Lim 알고리즘을 통해
스펙트로그램을 음성파일로
변환시켜주는 부분입니다.

딥러닝 학습

Attention



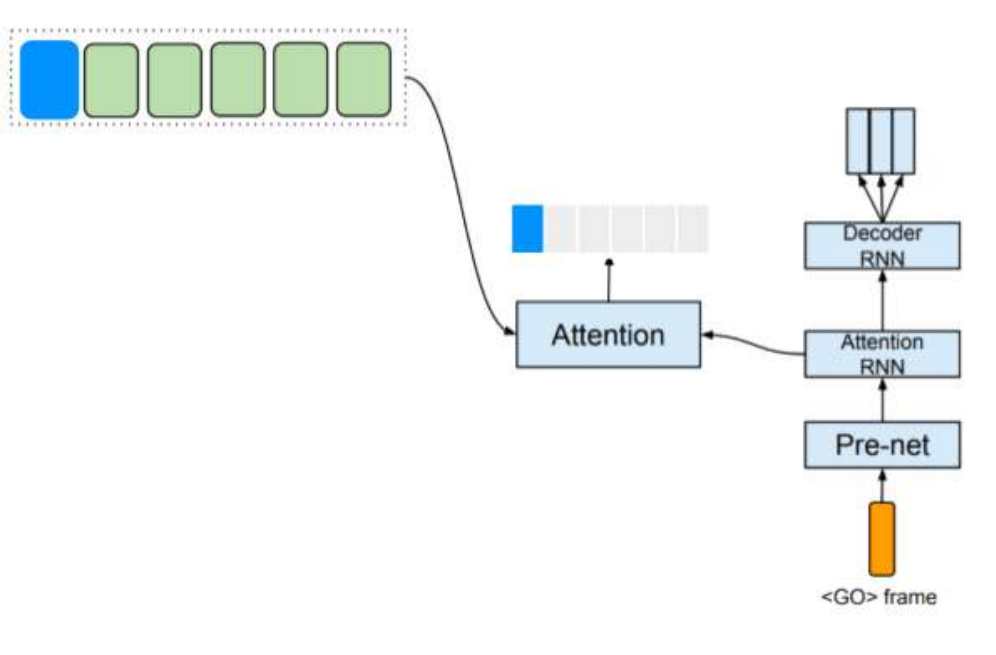
Attention은 간단히 말하면 음성을 만들 때 어디에 집중할 것 인가를 정해주는 부분입니다.

Attention이 중요한 이유는 일반화를 위해서 입니다.

즉, 학습하지 않았던 문장을 자연스럽게 말하기 위한 것이라고 할 수 있습니다.

딥러닝 학습

Attention

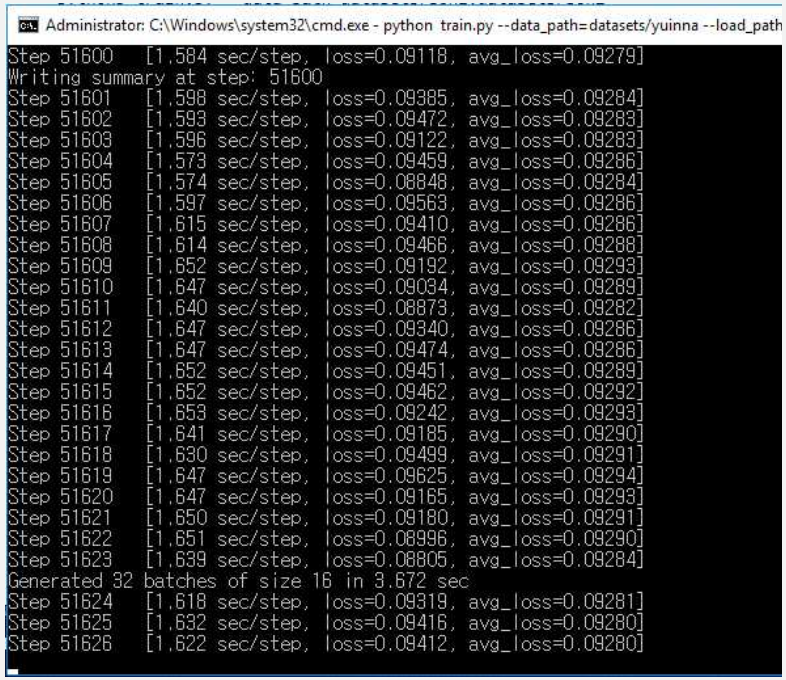


어디에 집중 할 것인지를 계산하고 스펙트로그램을 만드는 RNN에 Attention을 전달합니다.

여기까지가 Tacotron에서 딥러닝이 일어나는 과정입니다.

딥러닝 학습

실제 학습



실제로 달람 서버에서 학습을 돌리고 있는 화면입니다.

한 스텝당 약 1.6초가 소요되며
한 목소리 모델 당 약 10만 스텝 이상 러닝을 진행하였습니다.

러닝을 시키는 조건 중 하이퍼 파라미터를 조절하는 부분이 있었는데, 가장 최적화 되는 파라미터를 찾아서 사용하게 되었습니다.

또한, 수집한 데이터의 양이 적거나 데이터의 질이 떨어져서 러닝이 잘 되지 않을 때는 데이터를 더 수집하거나 음질이나 발음이 더 깨끗한 데이터를 수집하여 러닝을 다시 시작하여 업그레이드 하였습니다.

04 자기 평가

자기 평가

1. 기존의 TTS 방식보다 자연스러운 음성으로 재생되는가?

기존의 TTS 방식은 인간 개개인의 본연한 말투나 억양, 성조 등을 반영하지 못하고 그저 주어진 문장을 기계적으로 읽는 것에 비해 달람의 TTS 음성은 학습한 유명인의 말투 스타일을 비슷하게 재연했습니다.

하지만 네이버 오디오북 등에서 출시한 굉장히 자연스러운 인공지능 음성까지는 따라가지 못합니다. 입력하는 모든 텍스트를 확실하게 발음할 수 있는 수준까지는 아직 부족합니다. 그런 이유로 알람에 필요한 텍스트를 선별해 오버피팅하는 작업을 진행했습니다.

자기 평가

2. 선택한 연예인의 실제 음성과 비슷하게 학습되었는가?

두 팀원의 목소리를 각자 10시간 이상 학습하고,
완성된 모델에 연예인 음성을 3시간씩 학습시키는 작업을 진행했습니다.
결과물은 실제 연예인 음성과 비슷하게 재연되었습니다.

자기 평가

3. 특정 텍스트의 알람이 빠른 시간 내에 생성되는가?

알람 추가 화면에서 특정 텍스트를 입력하여 알람을 설정하면
약 30초 이내에 음성이 생성되므로
충분히 빠른 시간 내에 기능을 사용할 수 있습니다.

05 시연 영상

시연 영상

최종평가발표-시연동영상

https://github.com/kookmin-sw/2018-cap1-3/tree/master/doc/%EC%B5%9C%EC%A2%85%ED%8F%89%EA%B0%80%EB%B0%9C%ED%91%9C

에 올린 최종평가발표-시연동영상.mp4 파일을 보시면 되겠습니다.

(유튜브 링크: https://youtu.be/jDyjhYYQPxU)

감사합니다.