




쿠빅 침착맨 초대식



침착맨, 그가 걸어온 40년간의 여정

본명: 이병건 (李秉鍵)

출생: 1983년 12월 5일

국적: 대한민국

직업: 前 만화가, 現 스트리머

이력

- 육군 병장 만기전역
- 모범납세자 노원세무서장표창
- 건국대학교 시각디자인학 학사
- 한국인이 가장 사랑하는 유튜버



23년 3월 침착맨 오락가락 침가락 선언

침착맨, 무한으로 즐길 순 없을까..?



"정신적으로 오락가락"



팬들의 염려·응원 글



인기 웹툰작가 이말년, 개인방송 중단 선언



그래서 우리는 침착맨을
온라인에 가두기로 했다.

Contents

I. Overview

II. AI Chat Bot

III. Text to Speech

IV. Deployment

V. Conclusion

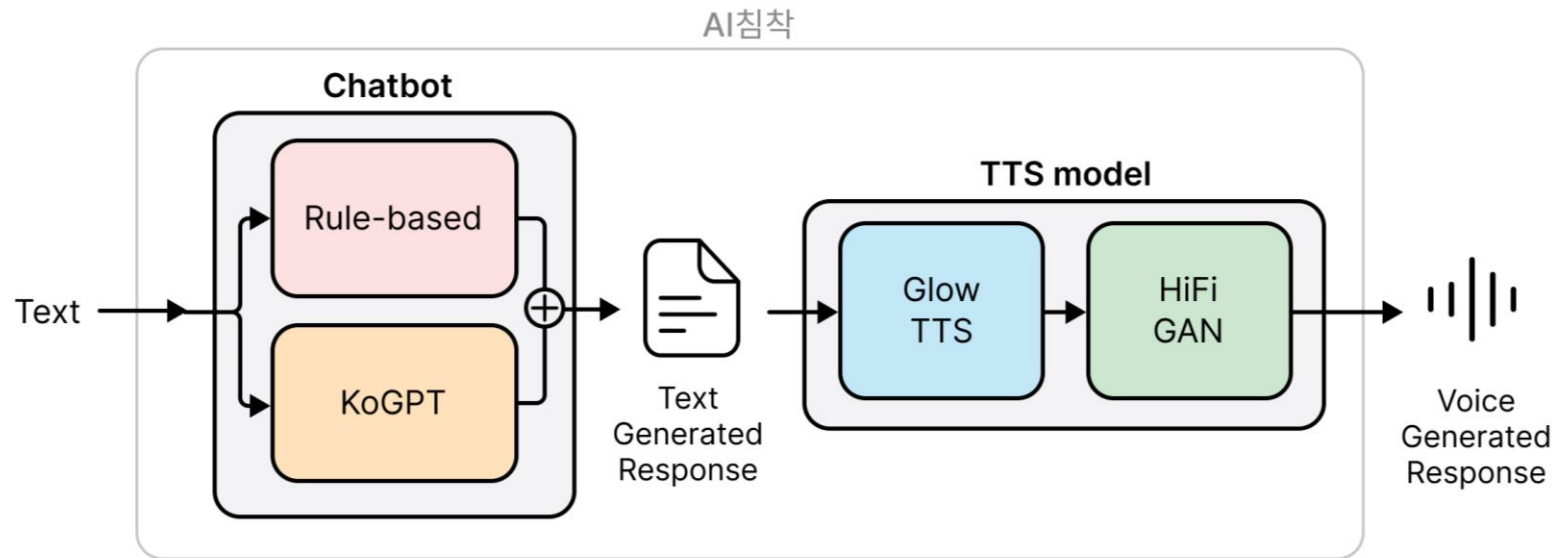
I. Overview

1. Overview

■ 프로젝트 파이프라인

음성봇 구현

1. 질문 text 입력
2. Chatbot으로 답변 생성
3. Text to Speech를 통해 text를 voice로 변환



1. Overview

■ 프로젝트 수행

프로젝트 구조

- 6 directories, 23 files로 구성
- GPU : Tesla V100 32 GB
- chatbot / common / voice로 구분하여 작업
- 용량이 큰 파일과 민감한 정보는 구글 드라이브 활용

개발환경



```
root@75728b3f1a92:~/Voice-Chatbot-Project/code# tree -L 2
.
|-- README.md
|-- chatbot_code
|   |-- ChatBotData.csv
|   |-- bot.py
|   |-- chatbot_last.csv
|   |-- inference.py
|   |-- inference.sh
|   |-- nohup.out
|   |-- output
|   |-- train.py
|   |-- train.sh
|   `-- txt
|-- common_code
|   |-- app.py
|   |-- csv_to_txt.py
|   |-- preprocessing.py
|   `-- speech_to_text.py
|-- ku_chim.jpg
|-- nohup.out
|-- run.py
`-- voice_code
    |-- alignment.py
    |-- aud_split.sh
    |-- infer_V2_.ipynb
    |-- inference.py
    |-- silence_slicing.py
    |-- train_glowtts.ipynb
    |-- train_hifigan.ipynb
    `-- tts-env

6 directories, 23 files
```


II. AI Chatbot

2. AI Chatbot

■ Data Collection

- 데이터 수집 방식
 - 유튜브의 영상 mp3 음성을 CLOVA Note로 텍스트 변환
- 데이터 종류
 - 왕십리로 날아온 편지 (46개)
 - 침착맨이 혼자 진행했던 코너 → 말투 수집 용이
 - 다른 코너에 비해 다양하고 일상적인 소재가 다수 포함



[왕십리로 날아온 편지 플레이리스트]

- 데이터 상세
 - 데이터 총 row 수 : 14496 row
 - 데이터 평균 길이 : 316 row/회, 19 row/사연

1	어떤 스트리머가 늦게 와서 너무 열받아요
2	어쩌면 좋죠
3	그런 고민이 있으셨군요
4	그러니까 이제 약속이라는 것은 진짜
5	무조건 지켜야 되는 서로 간에
6	중요한 건데 그거를 이제 남의 시간은 시간이라고 생각하지 않고 자기 시간만 시간이라고 생각하니까 그런 거죠
7	특히나 이제 스트리머 같은 경우는 일대일 약속하고는 더 상황이 좀 다른 게 기다리시는 분들이 천 명이면
8	천 명에 10초만 빠르면 얼마예요 만조죠 만조
9	만조를 뺐는 거예요 내 10초 늦으면 남의 만조를 뺐는 거야
10	그렇게 생각을 해서 굉장히 그거는 이렇게 지탄받아야 되는 행동이고 그런 분이 있으면 언제든지 제보해 주세요 제가 따끔하게 일침 놓겠습니다

[왕십리로 날아온 편지 스크립트]

2. AI Chatbot

■ Pre-processing

1

맞춤법 교정

Py-hanspell 사용

2

단어 통일

침착맨, 킹받네,
주펄, 침투부

3

사연 별 분할

기준
"다음 사연입니다"

4

반복 제거

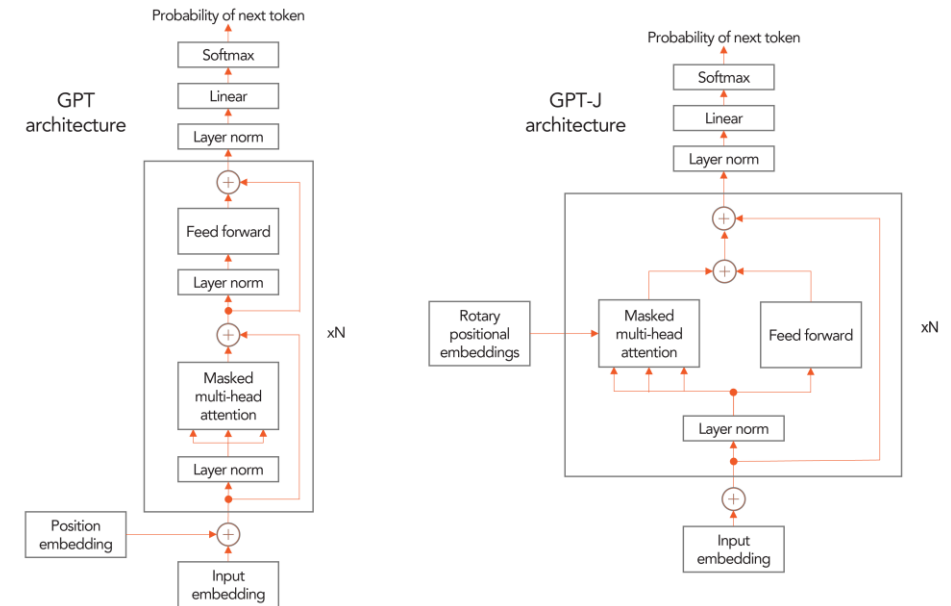
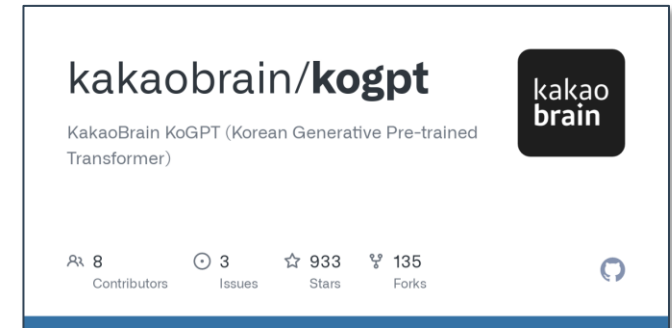
광고,
여기까지 하겠습니다,
물러나 보겠습니다

2. AI Chatbot

■ Ko-GPT

1. 모델 소개

- Kakao Brain의 KoGPT
- 한국어 버전 GPT3 (GPT-J)
- 60억개의 매개변수와 2000억개 토큰(token)의 한국어 데이터를 바탕으로 구축
- 다음과 같은 task 수행 가능
 - 1) 질문 답변
 - 2) 문장 요약
 - 3) 결론 예측
 - 4) 긍정/부정 판단

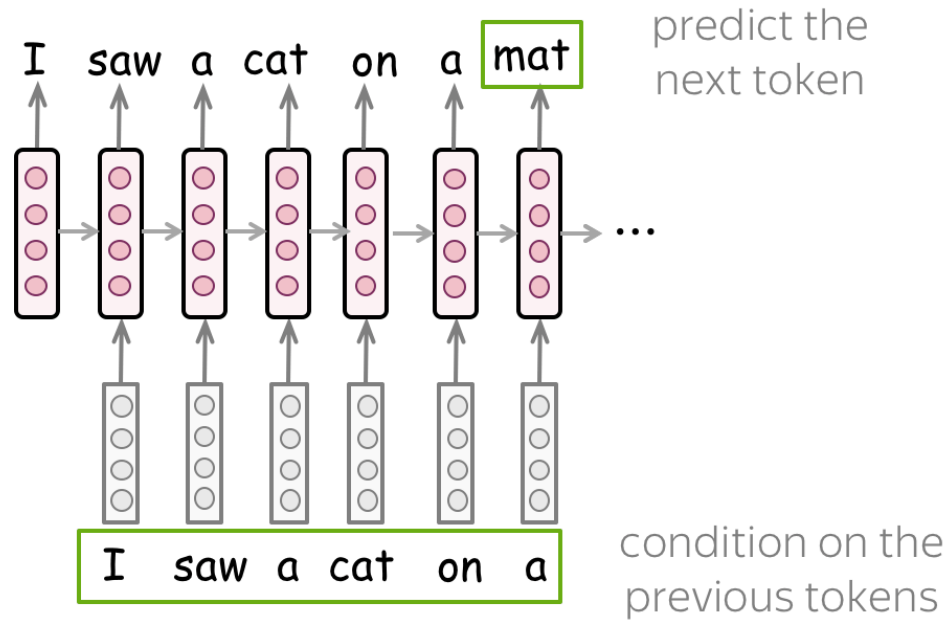


2. AI Chatbot

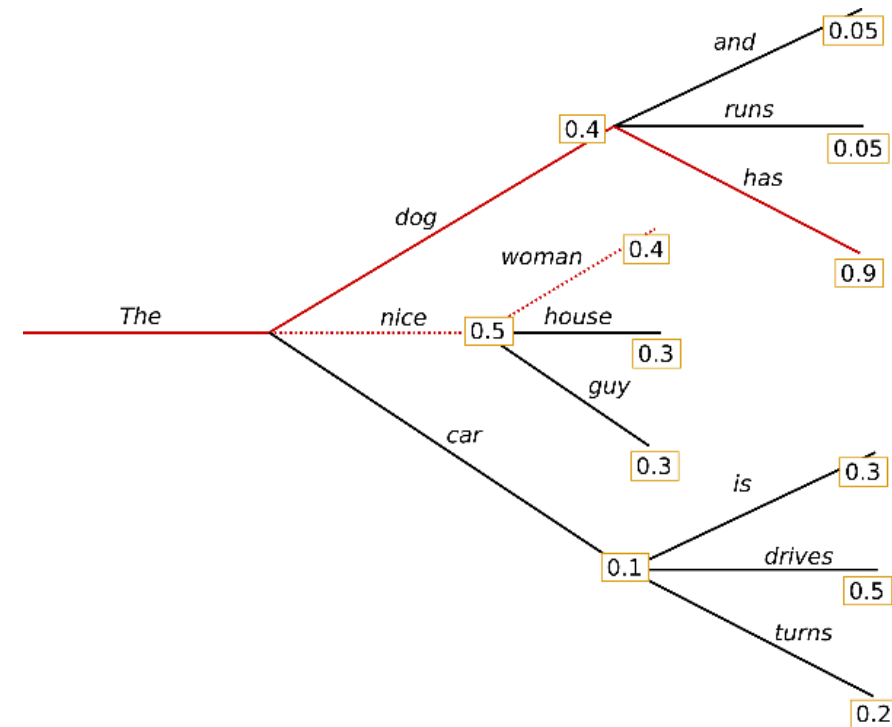
■ Ko-GPT

2. 모델 구조

Train(Next Token Prediction)



Inference(Beam Search)



2. AI Chatbot

■ AI Chatbot Model

1. 침착맨 라디오

침착맨이 이야기를 들려주는 것, 즉 생성이 핵심
⇒ 생성에 능한 GPT를 사용

GPT3의 생성 능력을 활용하여 침착맨 라디오는
흥미로운 이야기를 들려줍니다.

인공신경망 기반 챗봇



- 대량의 실제 대화 데이터를 통해 “인공신경망”을 학습하는 방식
- 생성 모델을 학습하여 복잡한 질문에 대해 좋은 품질의 답변을 생성할 수 있다.

2. AI Chatbot

■ AI Chatbot Model

2. 실시간 침착맨 챗봇

문제점

- 1) **제한된 입력 크기**
gpt3가 처리할 수 있는 프롬프트가
몇 문장보다 길 수 없음
- 2) **부정확한 생성 결과**
고유명사 문제, 부족한 데이터로 학습하여
조사와 의존명사만을 반복적으로 생성



유사도 기준 검색기반 챗봇을 채택

작동 방식

- 1) 사용자의 쿼리를 임베딩
- 2) Chatbot_data_for_Korean v1.0 데이터셋에서
쿼리와 벡터유사도가 높은 질문 검색
- 3) 선택된 질문에 해당하는 응답 출력

2. AI Chatbot

■ AI Chatbot Model

2. 실시간 침착맨 챗봇

침착맨과 실시간으로 대화를 주고 받는 것이
핵심

⇒ 답변 속도가 빠른 유사도 기준 검색 사용

Rule-based 시스템을 통해 실시간 침착맨 챗봇은 침착맨과의 자연스러운 대화 경험을 제공합니다.

규칙 기반 챗봇



- 질문에 포함된 의도를 파악한 후 의도에 대한 답변을
"유사도 분석과 검색"을 통해 찾아내는 방식
- 구현이 간단하며 많이 묻는 질문에 신속히 대응 가능하다.

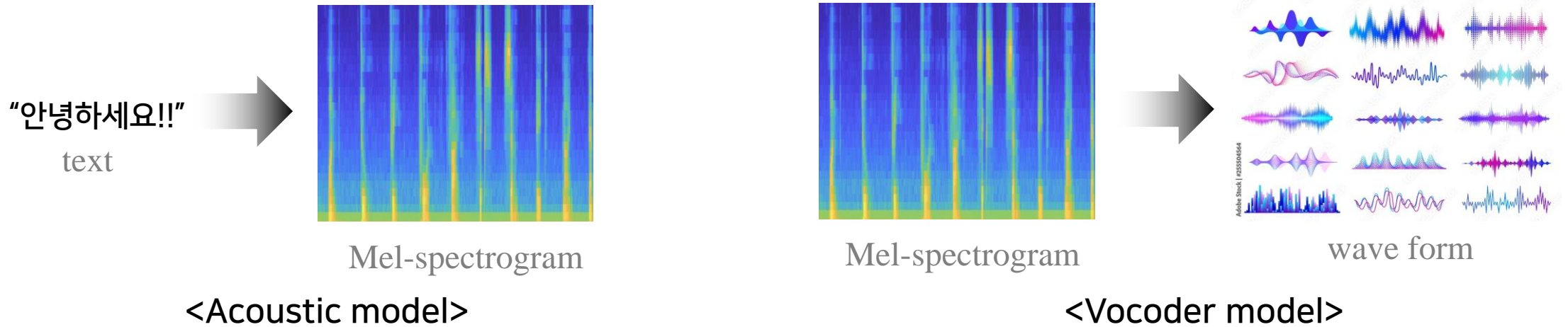
III. Text to Speech

3. Text to Speech

■ TTS 모델 구조

두 단계로 구성

1. Acoustic model (Text to Mel-Spectrogram)
2. Vocoder model (Mel-Spectrogram to Wave Form)

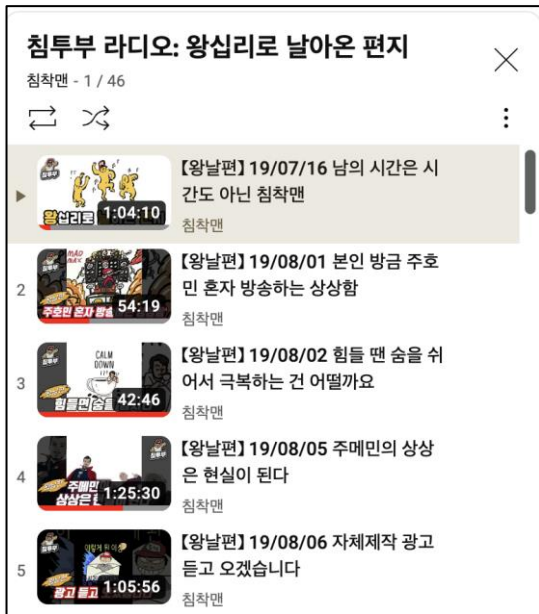


Acoustic model = Glow-TTS
Vocoder model = Hifi-GAN

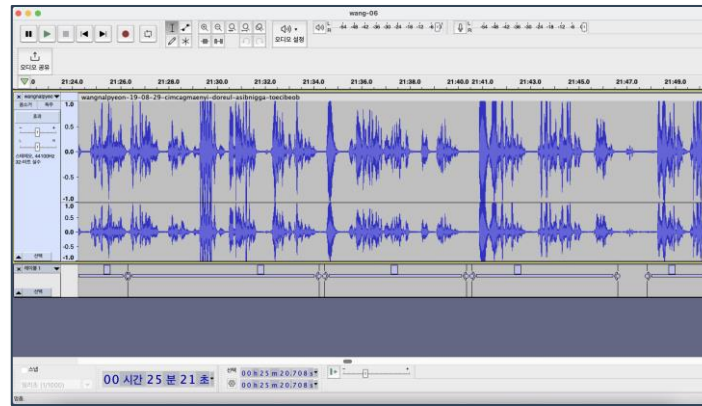
3. Text to Speech

■ 데이터 수집

- Youtube 침착맨 라디오 wav & text 수집
- Audacity program을 통해 5~12초 간격으로 편집
- 각 wav파일마다 대응되는 문장을 align해 저장

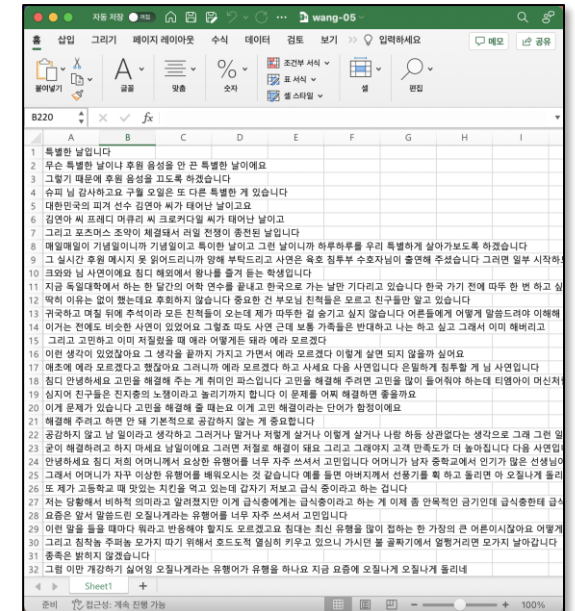


<collect wav & text>



<split wav data>

+



<split text data>

3. Text to Speech

■ Glow-TTS

1. 모델 소개

- "Generative Flow for Text-to-Speech via Monotonic Alignment Search" 논문으로 발표된 모델
- "Glow"는 "Generative Flow"의 약자로, 이 모델이 generative flow를 사용하여 음성 합성되었음을 의미함
- Glow-TTS 모델은 Acoustic model로 입력은 text이며 출력은 Mel-spectrogram임

2. 모델 장점

- 외부 aligner를 필요로 하지 않음
- 긴 문장에 대해서도 robust한 TTS 제공
- Auto regressive 모델에 비해 10배 이상 빠른 속도 제공

높은 성능과 빠른 속도를 가지고 있어 Real time TTS 모델 제작에 적합!

3. Text to Speech

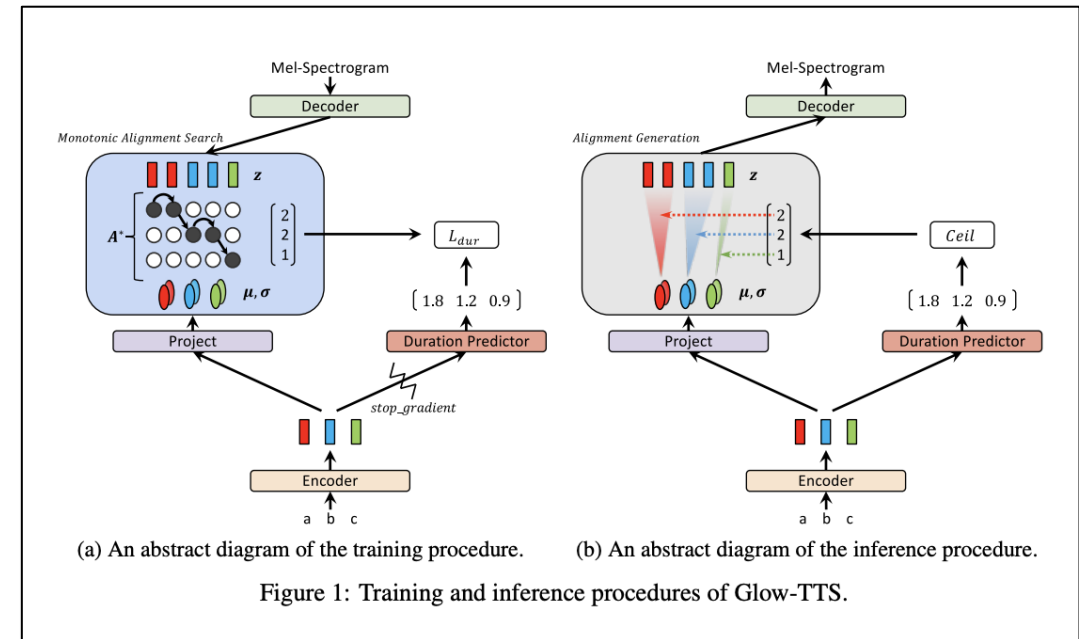
■ Glow-TTS

1. Monotonic Alignment Search(MAS)

- Encoder는 text를 받아 다중 가우시안 분포의 μ, σ 생성
- Decoder는 mel-spectrogram을 입력받고 latent representation을 생성한 후 두 값을 연결하는 분포를 학습

2. Flow-based model

- Encoder는 동일하게 수행하지만 Decoder는 Encoder에서 학습을 통해 만든 alignment 분포에 따라 만들어진 latent representation을 통해 Mel-spectrogram을 생성함
- 따라서 Decoder는 양방향으로 연산이 되는 Invertible한 형태를 가져야 함



$$\max_{\theta, A} L(\theta, A) = \max_{\theta, A} \log P_X(x|c; A, \theta)$$

3. Text to Speech

■ HiFi-GAN

1. 모델 소개

- HiFi-GAN의 이름은 "High Fidelity Generative Adversarial Networks"에서 유래됨
- "High Fidelity"는 고품질을 의미하며, "Generative Adversarial Networks"는 GAN을 의미.
- 따라서, HiFi-GAN은 고품질의 음성을 생성하는 GAN 모델을 뜻함

2. 모델 장점

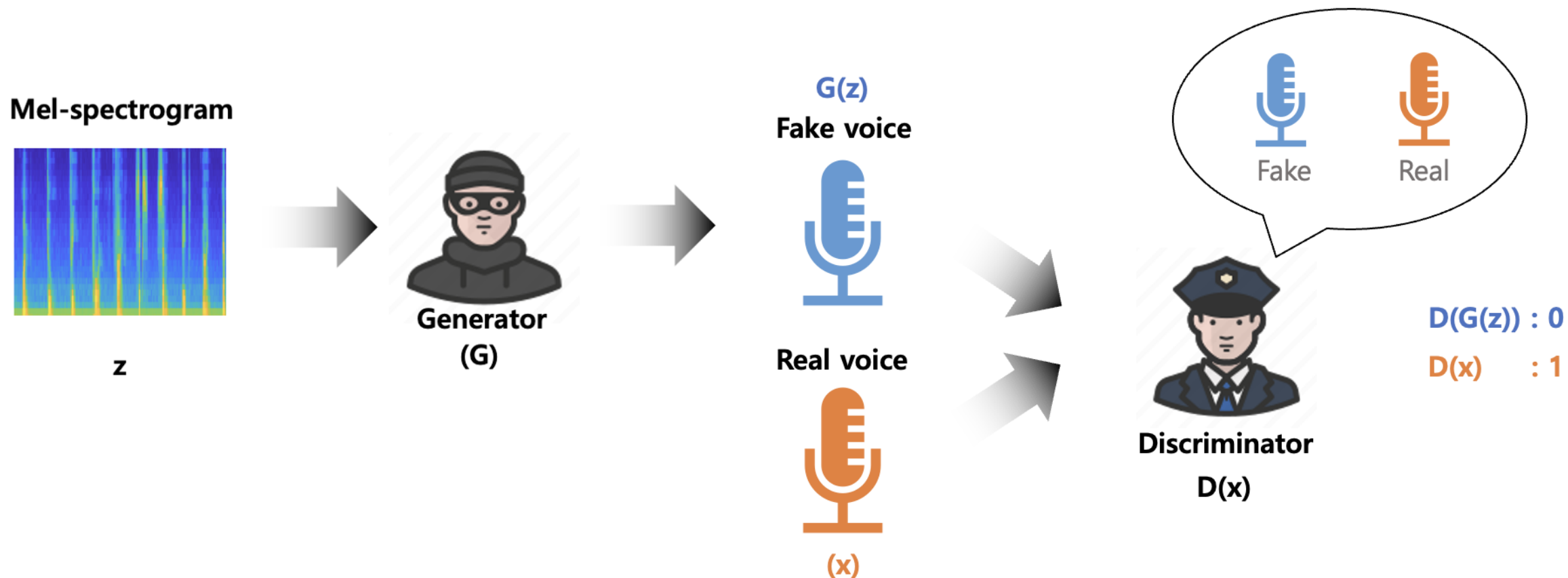
- GAN을 도입하여 더 높은 품질의 음성을 생성할 수 있음
- 노이즈가 있는 입력에서도 인간 수준의 음성을 생성 가능
- 기존 GAN Vocoder모델에 비해 생성자는 Residual block, 판별자는 sub-discriminator를 활용하면서 품질과 속도를 개선

작은 모델 크기와 빠른 속도, 높은 성능을 가지고 있어, 실시간 음성 생성에 적합!

3. Text to Speech

■ HiFi-GAN

3. 모델 학습 과정



3. Text to Speech

■ HiFi-GAN

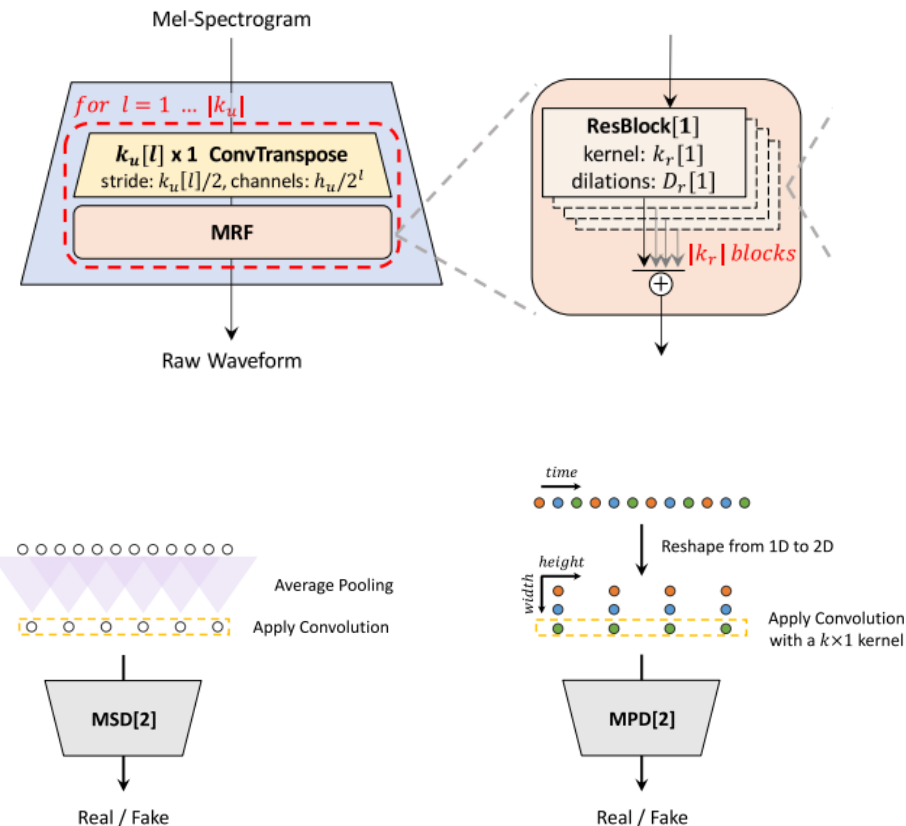
4. 모델 구조

1. Multi-Resolution Feature matching(MRF)

- 다양한 해상도(resolution)에서 생성자와 판별자의 특징(feature)을 비교하고 일치시키는 기법
- 해상도 변화에 Robust한 성격을 가지며 음질 향상을 가져옴

2. MSD & MPD

- Multi-Scale Discriminator(MSD), 전체 신호를 Average Pooling한 후 전반적 흐름 판별
- Multi-Period Discriminator(MPD), 전체 신호를 period별로 나누어 판별하며 구간적 흐름 판별






VI. Deployment

4. Deployment

■ Streamlit

쿠빅 침착맨 초대석

침착맨연KU소



반갑습니다 여러분의 귀염둥이 침착맨입니다.

실시간 침착맨 음성봇

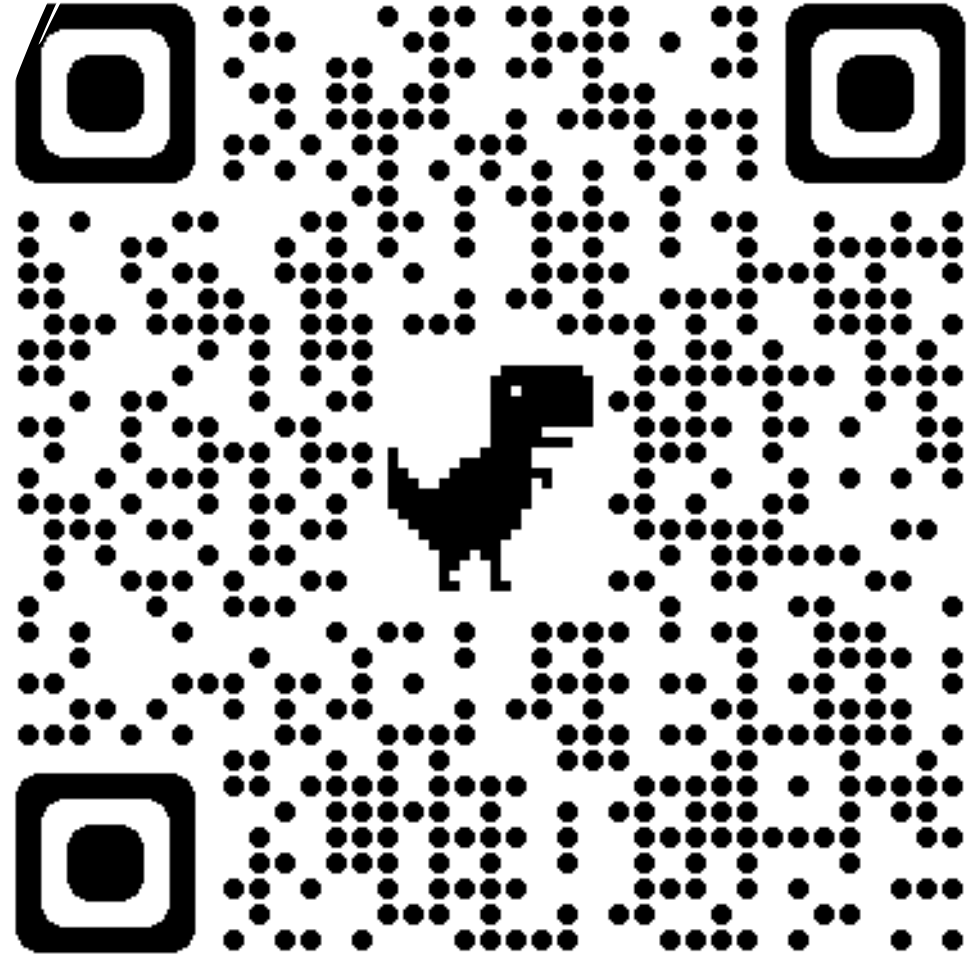
실시간 침착맨 음성봇

대화 보내기

침착맨 라디오

침착맨 라디오

<http://118.67.132.89:30007>



<Chrome Web Site QR Code>

V. Conclusion

5. Conclusion

■ 느낀점

1. 직접 수집한 데이터셋을 통한 프로젝트를 진행하며 데이터의 소중함을 깨달음
2. 멀티모달을 공부하며 다양한 분야를 이해하고 연결시키는 재미를 느낌
3. 모델을 단순히 만드는 것에서 그치지 않고 실제 사용자에게 전달되는 End-to-End 과정을 배움

■ 한계 및 후속연구

1. 양질의 데이터셋을 확보에 어려움 존재
2. 구어체와 고유명사가 있는 데이터셋을 사용 시 챗봇의 학습 성능이 저하됨
3. 더 많은 음성 데이터를 확보한다면 더 높은 성능을 기대할 수 있음
4. STT 구현하지 못한 것이 아쉬움

ESPIONAGE

딱 요정도만 하겠습니다